# Modelagem Estatística em Estudos Epidemiológicos. O Modelo Logístico.

Este exemplar corresponde à redação final da tese
devidamente corrigida e defendida por
Aluísio Jardim Dornellas de Barros
e aprovada pela comissão julgadora.

Prof. Dr. Euclydes Custódio de Lima Filho

Orientador

Campinas, 2 de março de 1990

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do grau de Mestre em Estatística.

Prof. Dr. Euclydes Custódio de Lima Filho

Orientador

## Pneumotórax

Febre, hemoptise, dispnéia e suores noturnos.
A vida inteira que poderia ter sido e que não foi.
Tosse, tosse, tosse.

Mandou chamar o médico:
— Diga trinta e três.
— Trinta e três ... trinta e três ... trinta e três ...
— Respire.

.....

— O senhor tem uma escavação no pulmão esquerdo e o pulmão direito infiltrado.
— Então, doutor, não é possível tentar o pneumotórax?
— Não. A única coisa a fazer é tocar um tango argentino.

*Manuel Bandeira*

# Agradecimentos

Em primeiro lugar quero agradecer ao Prof. Euclydes. Pela orientação da tese, sempre conseguindo colocar nas minhas certezas mais recentes germes de dúvida, não com o intuito de confundir, mas de provocar reflexão, questionamento das minhas "verdades" e conseqüentemente um conhecimento mais amadurecido e mais sólido. Pelo apoio que me deu nestes 4 anos, em todos os períodos, especialmente aqueles de crise nos quais às vezes eu duvidava da minha própria capacidade de resistir às pressões do curso. E especialmente, pela amizade, pelo exemplo profissional de dedicação, de ideal.

Quero também agradeder de maneira especial ao Prof. Antônio Carlos do Patrocínio que, com grandes doses de paciência e didática, me iniciou no estudo formal do Cálculo e da Álgebra Linear nos nossos seminários semanais, fazendo com que todos aqueles deltas e epsilons parecessem menos misteriosos.

Agradeço também a todos os docentes do departamento, que sempre me deram todo o apoio, seja durante os cursos, seja nas questões burocráticas, amenizando as dificuldades de um profissional da área de saúde que se aventurou a cursar um mestrado em Estatística. Quero mencionar explicitamente aqueles que lecionaram as disciplinas que cursei: Gabriela Stangenhaus, Ademir José Petenate, José Norberto W. Dachs, Sebastião de Amorim, José Antônio Cordeiro, José Ferreira de Carvalho, Eugênia Charnet e Regina Moran.

Ao Prof. Dr. Warren Winkelstein da Escola de Saúde Pública da Universidade da Califórnia, Berkeley, agradeço pelos dados utilizados no exemplo de aplicação, mediante autorização do Prof. Dr. Richard Brand, da mesma instituição.

Pelo apoio financeiro para o cumprimento de meu programa de mestrado, agradeço ao Conselho Nacional de Desenvolvimento Científico e Tec-

nológico (CNPq), à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e à Universidade Estadual de Campinas.

Aos funcionários do departamento, agradeço, pela atenção, pela simpatia. Em especial à Iara, que tanta dedicação tem pelo Laboratório de Estatística.

E, *last, but not least*, à Maristela, pela compreensão, pelo apoio, pela cumplicidade, pelo amor.

# Índice

# Lista de Tabelas

# Lista de Figuras

# Capítulo 1

# Introdução

Há séculos, a Medicina vem tentando dar solução aos inúmeros males que atingem o ser humano. Recentemente, uma nova área tem assumido um papel de importante colaboradora desse esforço: a Epidemiologia, que, ao contrário da pesquisa médica médica tradicional, voltada a um pequeno grupo de indivíduos, se preocupa com a ocorrência de doença dentro da população. Embora já em 1662 John Graunt tenha realizado um estudo considerado "epidemiológico", utilizando inclusive uma abordagem estatística para determinar taxas de mortalidade (Cornell, 1982 [4]), a estruturação da Epidemiologia como um corpo de conhecimento que pudesse caracterizá-la como uma ciência só ocorreu a partir deste século, mais precisamente a partir de grandes estudos epidemiológicos realizados durante e após a Segunda Guerra Mundial (Rothman, 1986 [19]).

Na opinião de Rothman, a Epidemiologia é ainda um embrião. Apesar de todo o crescimento que ela tem apresentado, vários fatos comprovam esta afirmação. O principal é que, ainda hoje, vários conceitos fundamentais da Epidemiologia são objeto de discussão e controvérsia, para não se falar na própria definição de Epidemiologia. Este é um fenômeno que não se observa nas ciências bem estabelecidas, onde a conceituação básica já é consenso. Grande parte das dificuldades, segundo este autor, são devidas a questões intrínsecas ao objeto do estudo epidemiológico, as populações humanas. A experimentação ocupa um lugar modesto na pesquisa epidemiológica e a maior parte da investigação é feita com base em estudos observacionais. Além disto, a ocorrência do evento em estudo é, em geral, rara. Estes fatores fazem com que a estimação de efeitos e a determinação de associações causais

sejam tarefas difíceis e sujeitas a uma série de elementos complicadores, como por exemplo, interação e confundimento (*confounding*), de definição também controversa.

Por outro lado, a Estatística tem sido uma colaboradora importante da Epidemiologia desde há muito tempo. William Farr, encarregado da estatística médica da Inglaterra e País de Gales durante 40 anos, a partir de 1839, fez com que a aplicação de métodos estatísticos passasse a fazer parte da rotina da investigação epidemiológica (Cornell, 1982 [4]). O grande avanço que a Estatística conheceu a partir da década de 1920, com os trabalhos de Fisher, contribuiu para que a colaboração entre estatísticos e epidemiologistas fosse cada vez maior. Citando novamente Rothman (1986), nem sempre este intercâmbio resultou positivo, visto que muitas vezes as técnicas estatísticas, desenvolvidas a princípio para outros campos de investigação, foram utilizadas sem muita crítica nos estudos epidemiológicos. No entanto é inegável que a Epidemiologia também se beneficiou muito da parceria. Muitas técnicas foram adicionadas ao seu arsenal de ferramentas e muitos estatísticos, com o passar do tempo, começaram a se dedicar exclusivamente à investigação epidemiológica, adaptando com critério as técnicas existentes e trabalhando na criação de ferramentas específicas. As técnicas de análise de dados categóricos, por exemplo, têm se desenvolvido muito devido à demanda da pesquisa médica. Referências obrigatórias nesse campo são o artigo de Grizzle, Starmer e Koch, de 1969 [7] e o livro de Bishop, Fienberg e Holland [2], publicado em 1975.

Na verdade, a Epidemiologia e a Estatística são aliadas naturais. O objetivo básico da Epidemiologia que é a *quantificação* do fenômeno doença em vários aspectos, faz com que o ferramental estatístico seja fundamental para o planejamento e análise dos estudos epidemiológicos. Desta forma, a formação em estatística é hoje fundamental para quem queira se intitular epidemiologista, embora não seja suficiente. E o fenômeno que observamos hoje é a aproximação que ocorre em muitas universidades entre os grupos de estatística e epidemiologia. Com muita freqüência encontramos atualmente epidemiólogos e bioestatísticos trabalhando num mesmo departamento ou em departamentos distintos, mas dentro de uma mesma faculdade de Medicina ou de Saúde Pública.

De toda forma, a preocupação básica é a quantificação da ocorrência da doença nos seus vários aspectos, como Kleinbaum, Kupper e Morgenstern (1982 [10]) deixam claro na introdução de seu livro "Epidemiologic Re-

search". E é neste sentido que se desenvolve nosso trabalho. Buscamos detalhar técnicas que permitam que os dados epidemiológicos sejam trabalhados de forma quantitativa, que possam auxiliar o epidemiologista a extrair o máximo de informação dos seus dados. Gostaríamos também que a divulgação destas técnicas incentivasse o pesquisador a realizar estudos de cunho quantitativo, um passo adiante dos estudos qualitativos que apontam causas ou fatores de risco, ou ainda sugerem relações entre tais fatores, mas são incapazes de indicar em quanto tais ou quais fatores influem no agravo em estudo.

Mais especificamente, este trabalho aborda a modelagem estatística de estudos epidemiológicos, quando a resposta observada é binária, utilizando a função *logito*. Como detalharemos mais adiante, o **Modelo Logístico** é extremamente interessante para a modelagem de estudos prospectivos e em especial dos estudos retrospectivos. No Capítulo 2 abordamos os conceitos fundamentais da Epidemiologia, que são necessários ao desenvolvimento das idéias relacionadas à modelagem estatística. A literatura utilizada constituise basicamente de obras consagradas em Epidemiologia, como os trabalhos de Miettinen, Kleinbaum et alii, Breslow & Day e Rothman. No Capítulo 3 fazemos uma breve discussão sobre modelagem estatística dos estudos epidemiológicos e apresentamos o modelo logístico, suas principais características matemáticas, diversas possibilidades de aplicação em Epidemiologia e a justificativa para sua utilização com estudos de caso-controle. Este capítulo está baseado em Cox (1970 [5]) e em Kleinbaum et alii (1982 [9]) e a última seção em um artigo de Prentice & Pyke de 1979 [17]. No Capítulo 4 desenvolvemos um estudo mais técnico da verossimilhança do modelo, técnicas de estimação, estatísticas para testes de hipóteses relacionados aos parâmetro do modelo e introduzimos as idéias básicas para diagnóstico no modelo logístico. Além das obras já citadas são utilizados os trabalhos de Bishop et alii (1975 [2]), Fienberg (1977 [6]) e Pregibon (1981 [16]). Por fim, no Capítulo 5, trabalhamos um exemplo, de forma a demonstrar a aplicabilidade das técnicas apresentadas e facilitar o trabalho daqueles que, eventualmente, queiram se servir deste volume para aplicar esta modelagem a seus conjuntos de dados.

# Capítulo 2

# Epidemiologia – conceitos fundamentais

## 2.1 Definição e objetivos

A Epidemiologia, que em seu início cuidava de dar suporte metodológico ao estudo e controle de epidemias, tem hoje um campo de ação bastante vasto, que se estende às doenças infecciosas ou não, agudas e crônicas. Mostra também uma grande preocupação com os serviços de atenção à saúde, no tocante à sua avaliação e à avaliação do impacto que medidas tomadas por estes serviços têm sobre a saúde da população (Kleinbaum et alii, 1982 [10]).

A melhor definição de Epidemiologia, segundo Rothman (1986 [19]), é aquela atribuída a Gaylord Anderson: *o estudo da ocorrência da doença.* Outras ciências podem se ocupar do estudo da doença, interessadas em suas manifestações clínicas ou métodos de cura, mas, a Epidemiologia se ocupa da *ocorrência* da doença.

Este termo chave implica que o alvo da atenção da Epidemiologia é a população. Afinal, só se pode observar e analisar a ocorrência de uma doença em um contexto populacional, tanto que uma outra definição corrente de Epidemiologia é: *o estudo da doença em relação às populações* (Rose & Barker, 1979 [18]). Assim, ao contrário da Medicina clínica, que se preocupa com o indivíduo, a Epidemiologia se debruçará sobre as populações para estudar a ocorrência de doença com o objetivo de:

1. Determinar padrões de ocorrência das doenças segundo variáveis geográficas, sociais, econômicas, antropométricas, etc., com a intenção básica de traçar um perfil da situação de saúde da população.

2. Determinar padrões de ocorrência das doenças segundo fatores considerados "de risco", com o intuito de estabelecer quais deles efetivamente influenciam seu aparecimento e quantificar esta influência, elucidando sua *etiologia.*

3. Predizer a ocorrência das doenças na população, possibilitando

4. Controlar a ocorrência das doenças pela ação preventiva e pela orientação das ações curativas (Kleinbaum et alii, 1982 [10]).

Dois termos foram usados freqüentemente na definição de Epidemiologia e na apresentação dos seus objetivos: *população* e *doença.* É importante que se discuta como o epidemiologista os entende.

A pesquisa epidemiológica é sempre voltada para um determinado grupo humano. Este grupo pode ser amplo, abrangendo toda a população de um país, por exemplo, quando se pesquisa a influência do hábito alimentar em afecções cardio-circulatórias. Mas também pode ser muito restrito, como no caso de se estar interessado no efeito da inalação de partículas em trabalhadores de indústrias cerâmicas. Assim, podemos definir **população** de uma forma concisa como sendo um determinado grupo, incluindo tanto elementos sadios quanto doentes, sobre cuja saúde se pretende fazer alguma afirmação (Rose & Barker, 1979 [18]). Na verdade, este grupo, ou *população alvo*, não será utilizado para o estudo e sim um subconjunto dele, que se convencionou chamar de *população em estudo.* É desta população que se vai retirar uma amostra para que uma inferência estatística possa ser feita.

A escolha da *população alvo* é feita em função do interesse do pesquisador. Sua definição pode ser de caráter geográfico, como moradores de uma determinada cidade; ocupacional, como trabalhadores exercendo determinada atividade; etário, como todas as crianças abaixo de 2 anos de idade, etc. O fundamental é que a definição seja precisa e permita que se saiba com clareza quais indivíduos pertencem a esta população.

Por outro lado, **doença** deve ser entendida de uma forma ampla. Em geral, o termo doença designa uma afecção que provoca algum tipo de dano ao indivíduo, sendo a definição estrita do termo bastante controversa.

No contexto deste trabalho entendemos doença como qualquer manifestação encontrável na população que seja de interesse do pesquisador. Assim, pode-se estar interessado na ocorrência de sarampo, mas, também pode-se querer estudar, por exemplo, o estado de imunidade de uma população após um programa de vacinação em massa. Ou seja, embora estejamos usando o termo doença, este tanto pode ter o sentido tradicional, como pode ser entendido como uma outra manifestação qualquer, relacionada à saúde do indivíduo, que seja de interesse.

Uma outra característica da Epidemiologia é, visto que se busca entender como uma doença ocorre em uma população, a necessidade de estudar tanto os indivíduos doentes como os sãos. A medida de ocorrência de doença mais corriqueira é uma razão entre o número de doentes observados e a soma do número de doentes e não doentes. Desta forma, para a Epidemiologia, os indivíduos livres do agravo em questão são tão importantes quanto os doentes.

## 2.2 Medidas de freqüência das doenças

Para alcançar os objetivos da Epidemiologia fazem-se necessárias medidas que quantifiquem a ocorrência de uma doença na população. Existem muitas medidas que são utilizadas num contexto geral ou em situações específicas. Abordaremos aqui apenas as 3 fundamentais, das quais as outras são, de alguma forma, derivadas: a *incidência cumulativa*, a *densidade de incidência* e a *prevalência*.

### 2.2.1 Incidência Cumulativa

Esta é a medida mais comumente usada na descrição da freqüência de uma doença. Considere uma determinada população fixa, isto é, uma população em que não há entrada de novos indivíduos durante a observação e da qual só saem indivíduos acometidos pelo agravo de interesse. Esta população é observada de um instante $t_0$ a um instante $t_1$. No instante $t_0$ esta população tem $N_0$ indivíduos sãos. Define-se como incidência cumulativa a razão entre o número de casos novos de doença (D) observados neste intervalo de tempo ($\Delta t = t_1 - t_0$) e o tamanho da população no instante inicial de observação. Esta medida é uma proporção e pode ser interpretada como uma

estimativa do *risco*, isto é, da probabilidade de se adoecer neste intervalo de tempo (Morgenstern et alii, 1980 [14]):

$$\hat{R}_{\Delta t} = IC_{\Delta t} = \frac{D}{N_0} . \tag{2.1}$$

A incidência cumulativa varia de 0 a 1, é adimensional e depende intrinsecamente do tempo de observação $\Delta t$. Quanto maior este tempo, maior deve ser a proporção observada, e vice-versa, pois varia-se o tempo em que as pessoas estiveram sob risco. Por isso, a medida é apresentada indexada por $\Delta t$ e quando usada na prática deve ser sempre acompanhada da especificação do tempo de observação.

Esta medida, apesar de simples, apresenta algumas dificuldades ao ser aplicada em situações reais. A suposição de população fixa implica que todos os indivíduos que estavam presentes no início da observação ficaram até o final do período ou saíram por terem adquirido a doença em questão e que nenhum novo indivíduo entrou na população observada durante o intervalo. Em primeiro lugar, não é difícil que indivíduos adoeçam ou morram por outras causas que não a de interesse durante a observação, especialmente se este período se torna longo. De uma forma bastante rígida, um indivíduo não deveria ser incluído no denominador a não ser que tivesse sido observado por todo o período ou que tivesse adoecido pela causa de interesse. Se alguém sai do estudo antes de seu término por outras causas, não se sabe se este indivíduo desenvolveria ou não a doença no tempo restante de observação.

Além disso, é comum em estudos prolongados que os indivíduos sejam incluídos no grupo observado ao longo de um certo período, o que faz com que cada um participe do estudo durante um intervalo de tempo diferente. Neste caso, o uso da fórmula 2.1 para estimar o risco claramente não é satisfatório, visto que trata de maneira igual contribuições diferentes de cada indivíduo para o experimento. Assim é que se buscou uma outra medida de incidência que superasse estas deficiências: a densidade de incidência.

### 2.2.2 Densidade de Incidência

Apresentamos esta medida com a designação proposta por Miettinen (1976 [12]). Outros a chamam de *taxa de incidência* ou *força de morbidade*. Ela é definida como a razão entre o número de casos novos (D) observados durante o período de estudo e o tempo total de observação constituído pela

soma dos tempos com que cada indivíduo contribuiu para a pesquisa. Se temos $N$ indivíduos que foram observados durante o estudo, cada um tendo sido observado por um tempo $t_i$, temos que a densidade de incidência é

$$DI = \frac{D}{\sum_{i=1}^{N} t_i} . \tag{2.2}$$

A densidade de incidência deve ser estudada com cuidado, visto que não é uma medida com características usuais – ela não é uma proporção e não dá idéia de risco. Seu numerador é constituído por uma freqüência e seu denominador por uma medida de tempo, de sorte que sua dimensionalidade é 1/tempo ($t^{-1}$). A interpretação não é imediata, mas uma comparação com velocidade é pertinente. A densidade de incidência pode ser vista como uma "velocidade média" de ocorrência da doença. Quanto maior esta "velocidade", maior será a força de morbidade da doença.

Para deixar claro que o tempo com que se trabalha nesta medida não é o tempo ordinário, mas o tempo vivido por um grupo de pessoas observadas, este denominador costuma ser referido como uma medida de *tempo-pessoa*[1]. Uma determinada quantia de tempo-pessoa pode ser obtida de várias formas. Dez anos-pessoa podem ser obtidos pela observação de 10 pessoas por 1 ano ou pela observação de 20 pessoas por 0,5 ano. Ou seja, não importa quanto tempo se leva para realizar o estudo, mas quanto tempo de experiência individual se observou. Desta forma, a densidade de incidência traz a noção de temporalidade embutida e não há necessidade de se fazer referência ao período de estudo ao se apresentar a medida como resultado de uma investigação.

O intervalo de variação desta medida é de 0 a infinito. Para que a densidade de incidência seja 0, basta que não se observe nenhum caso de doença durante o estudo. Conforme se observe um número crescente de casos em intervalos curtos de tempo o valor observado crescerá, chegando a infinito numa situação teórica de grande número de casos em um intervalo extremamente curto de observação (e.g., uma explosão atômica). Lembrar do comportamento da função $1/t$ para $t$ não negativo facilita a compreensão desta questão.

[1]Preferimos este, ao invés do termo mais freqüente em português, pessoa-tempo, por acreditarmos que, além de ser uma tradução melhor do termo original *person-time*, transmite com mais clareza o conceito envolvido.

A diferença entre a densidade de incidência e a incidência cumulativa está no denominador. Em vez de se usar uma contagem dos indivíduos da população, a densidade de incidência usa tempo. Teoricamente, esta mudança resolve os problemas apresentados anteriormente. Não há mais dificuldade no tratamento de indivíduos que entraram depois do início da observação ou saíram antes, pois o que será levado em conta é o tempo com que cada elemento contribui para o total de tempo-pessoa de observação. Num estudo prospectivo, onde se tenha controle total dos indivíduos que entram e saem do grupo observado, não há dificuldade em se obter a medida total de tempo-pessoa. Como indicado em 2.2, basta se somar a contribuição individual de tempo para cada elemento que participou do estudo.

Existem outras situações, no entanto, em que a obtenção do denominador não é tão simples. Se nosso interesse recai sobre uma população muito grande, em que o controle preciso é impossível, teremos que encontrar meios para aproximar o total de tempo-pessoa. Conceitualmente, podemos imaginar a experiência de tempo-pessoa de 2 tipos distintos de população: uma população *fixa* e uma população *dinâmica* (Rothman, 1986 [19]).

População fixa, como já comentamos, é aquela em que não há entrada de novos indivíduos durante a observação e só ocorre a saída deles por acometimento pelo agravo de interesse. Neste caso, o cálculo do total de tempo-pessoa se faz de maneira similar ao caso apresentado anteriormente. Visto que a contribuição individual de tempo equivale ao intervalo que vai do início do estudo até o momento do aparecimento da doença no elemento, em geral não se tem dificuldade de determiná-la de forma exata.

Contrariamente, numa população dinâmica, isto é, numa população em que os indivíduos entram e saem a qualquer momento, torna-se complicado manter um controle individual para um grupo grande. Acontece que na maioria dos casos reais, trabalhamos com populações dinâmicas, sujeitas a múltiplas causas de morbidade e mortalidade, a migrações e à natalidade. Na impossibilidade de se manter controle rígido sobre cada indivíduo, temos que lançar mão de meios que nos permitam aproximar o total de tempo-pessoa vivido pelo grupo durante o intervalo de interesse.

Uma das suposições que se faz com freqüência é de que a população de interesse seja *estável*. Isto significa que as forças que provocam entrada ou saída de indivíduos na população estão em equilíbrio, de forma que o número total de pessoas não se altera no decorrer da observação. Se esta suposição é aplicável, o cálculo do total de tempo-pessoa fica muito simplificado, sendo

apenas o produto do número de pessoas pelo tempo de observação e temos

$$DI = \frac{D}{N\,\Delta t} \tag{2.3}$$

onde $N$ é o tamanho da população e $\Delta t$ o tempo de observação.

### Análise de componentes seqüenciais

O acometimento ou a morte por uma determinada doença pode ser visto como resultado de uma *seqüência de eventos patogênicos.* Desta forma, é razoável supor que seja possível decompor taxas globais em parcelas referentes a estes eventos. O mais freqüente é subdividir o período em que o indivíduo está sob risco de contrair uma doença até sua morte (supondo que não há riscos competitivos) em 2 períodos. O primeiro vai da exposição até a ocorrência da doença e o segundo deste momento até a morte. Queremos, então, relacionar a taxa de mortalidade por esta causa com a densidade de incidência e com a taxa de letalidade da doença (Morrison, 1979 [15]).

Por taxa de mortalidade, estamos designando a densidade de incidência de mortes pela causa em questão. Taxa de letalidade é a densidade de incidência de mortes entre os doentes e chamamos simplesmente de densidade de incidência àquela da ocorrência da doença em questão.

Suponha uma população fixa que é acompanhada até a morte do último indivíduo. A densidade de incidência será:

$$DI = \frac{N}{\sum_{i=1}^{N} t_i} = \frac{N}{\frac{N}{N}\sum_{i=1}^{N} t_i} = \frac{N}{N\bar{T}} = \frac{1}{\bar{T}} \tag{2.4}$$

onde $\bar{T}$ é a média das contribuições individuais de tempo-pessoa.

Ou seja, densidade de incidência, neste caso também taxa de mortalidade, corresponde à inversa do tempo médio até a morte. Por exemplo, uma $DI = 0,04$/ano-pessoa corresponde a uma esperança de vida de 25 anos. Se, em vez de morte, nos detivermos no aparecimento da doença, esta relação continua válida, sob algumas condições. Em geral, exige-se que a população em estudo seja estável e que a densidade de incidência seja constante para cada grupo etário.

Dividindo o tempo até a morte em dois intervalos, um até o aparecimento de doença e outro até a morte podemos escrever $\bar{T} = \bar{T}_1 + \bar{T}_2$, sendo

$\bar{T}_1$ o tempo médio até o aparecimento de doença e $\bar{T}_2$ o tempo médio entre o adoecer e a morte. Pelo que já foi visto podemos escrever a taxa de mortalidade $M$ como

$$M = \frac{1}{\bar{T}} = \frac{1}{\bar{T}_1 + \bar{T}_2}. \tag{2.5}$$

Mas, acabamos de mostrar que $\bar{T}_1$ pode ser interpretado como o inverso da densidade de incidência, assim como $\bar{T}_2$ pode ser visto como o inverso da taxa de letalidade ($L$). Nossa relação final será

$$M = \frac{1}{\bar{T}_1 + \bar{T}_2} = \frac{1}{\frac{1}{DI} + \frac{1}{L}} \tag{2.6}$$

Concluímos que, conhecendo a duração média da doença e o tempo médio até adoecer, pode-se estimar a taxa de mortalidade por aquela causa, respeitadas as suposições colocadas acima. Mais do que isso, conhecemos a relação entre medidas importantes da doença.

### 2.2.3 Relação entre as medidas de incidência

É possível, também, relacionar a incidência cumulativa com a densidade de incidência. Considere $N_0$ o tamanho de uma população fixa de indivíduos sadios no instante inicial de observação $t_0$, e $N_t$ o tamanho dela no instante $t$. Neste intervalo $[t_0, t]$ a incidência cumulativa é

$$IC_t = \frac{N_0 - N_t}{N_0} = 1 - \frac{N_t}{N_0}. \tag{2.7}$$

A densidade de incidência no intervalo $[t, t+\Delta t]$ pode ser escrita como

$$DI_t \approx \frac{-\Delta N}{N_t \, \Delta t}, \tag{2.8}$$

onde $-\Delta N = -(N_t - N_0)$ é o número de casos novos no intervalo e $N_t \, \Delta t$ representa a área hachurada na Figura 2.1 e é uma aproximação da medida de tempo-pessoa do intervalo. Fazendo $\Delta t$ tender a zero, podemos escrever, usando a notação do cálculo diferencial,

$$DI_t = \frac{-dN}{N_t \, dt} \tag{2.9}$$

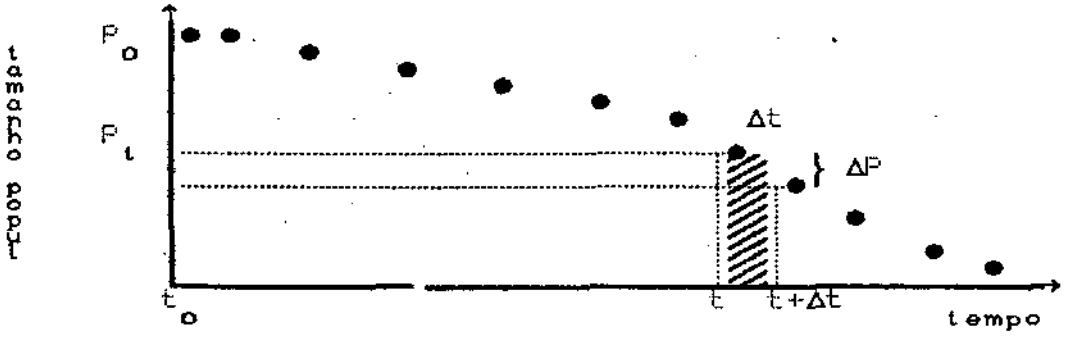


Figura 2.1: Comportamento de uma população fixa em relação a uma doença no decorrer do tempo.

$$-DI_t\,dt = \frac{dN}{N_t}\,. \tag{2.10}$$

Integrando ambos os lados de $t_0$ a $t$,

$$-\int_{t_0}^{t} DI_t\,dt = \int_{t_0}^{t} \frac{1}{N_t}\,dN = \log(N_t) - \log(N_0), \tag{2.11}$$

aplicando a função exponencial e usando a relação 2.7 temos

$$\exp\left(-\int_{t_0}^{t} DI_t\,dt\right) = \frac{N_t}{N_0}$$

$$IC_t = 1 - \exp\left(-\int_{0}^{t} DI_t\,dt\right)\,. \tag{2.12}$$

Esta relação (2.12) pode ser estimada por

$$IC_t = 1 - \exp\left(-\sum_{i=1}^{s} DI_i\,\Delta t_i\right), \tag{2.13}$$

se garantirmos que a soma dos intervalos $\Delta t_i$ resultam no intervalo $\Delta T = [t_0, t]$ e que a densidade de incidência em cada $\Delta t_i$ é constante.

Se a densidade de incidência é constante para todo o intervalo $\Delta T$, então a relação pode ser simplificada para

$$IC_t = 1 - \exp\left(-DI \sum_{i=1}^{s} \Delta t_i\right)$$

$$IC_t = 1 - \exp\left(-DI\,\Delta T\right). \tag{2.14}$$

A expansão em série de Taylor de $e^x$ é

$$e^x = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \cdots$$

Para um $x$ pequeno, $1 + x$ já é uma boa aproximação de $e^x$. Assim, substituindo em 2.13, temos

$$IC_t \approx 1 - \left(1 - \sum_{i=1}^{s} DI_i\,\Delta t_i\right) = \sum_{i=1}^{s} DI_i\,\Delta t_i\,. \tag{2.15}$$

Ou, para $DI$ constante em $\Delta T$,

$$IC_t \approx DI\,\Delta T\,. \tag{2.16}$$

Estas relações entre as medidas de incidência são importantes, na medida em que nem sempre é possível obter uma estimativa direta de uma delas. Assim, através do conhecimento de uma, pode-se estimar a outra.

### 2.2.4 Prevalência

Define-se prevalência como a proporção de uma população que está acometida por uma determinada doença em um certo instante no tempo. Ao contrário da incidência cumulativa, que é uma medida tomada em um intervalo de tempo, a prevalência é a proporção de doentes *em um determinado momento.* Por isso, também é chamada de *prevalência pontual.* A prevalência num tempo $t$ é dada por

$$P_t = \frac{C_t}{N_t}, \tag{2.17}$$

onde $C_t$ é o número de doentes na população no instante $t$ e $N_t$ o tamanho desta população.

Esta medida nos dá uma estimativa da probabilidade de um indivíduo estar doente naquela população e naquele instante. Como veremos adiante, a prevalência depende da densidade de incidência e da duração média da doença.

### 2.2.5 Prevalência e incidência

O conceito de população estável pode ser estendido para o comportamento da população em relação a uma doença. Assim, se o número de pessoas que adoece é o mesmo das pessoas que se curam, dizemos que a população é estável em relação à doença de interesse (Rothman, 1986 [19]). Supondo que isto seja verdade para uma determinada população, temos, usando (2.16) que o número de pessoas que adoece num determinado período de tempo é $DI\,\Delta t(N-C)$, onde $DI$ é a densidade de incidência da doença, $\Delta t$ o período de tempo, $N$ o número de pessoas da população e $C$ o número de doentes. Da mesma forma, o número de pessoas que se curam neste mesmo período será $DI'\,\Delta t\,C$, onde $DI'$ é a densidade de incidência de cura. Pela suposição de equilíbrio da população, o número de pessoas que adoecem é o mesmo das que se curam, o que equivale a

$$DI\;\Delta t\,(N-C) = DI'\,\Delta t\,C\;. \tag{2.18}$$

Usando relações já estudadas (2.4), podemos dizer que $DI' = 1/\bar{T}$, onde $\bar{T}$ é o tempo médio de duração da doença. Substituindo e rearranjando,

$$\frac{C}{(N-C)} = \frac{C/N}{1-(C/N)} = DI\,\bar{T}. \tag{2.19}$$

Ou seja, o **odds de prevalência** (*prevalence odds*) é igual à densidade de incidência multiplicada pela duração média da doença. Através desta relação fica fácil encontrar uma boa aproximação para a prevalência, no caso de doenças raras. Quando há poucos doentes na população, o odds de prevalência se aproxima da prevalência e temos

$$P = \frac{C}{N} \approx DI\,\bar{T}. \tag{2.20}$$

Determinando a relação exata entre prevalência e densidade de incidência encontramos que

$$P = \frac{C}{N} = \frac{DI\,\bar{T}}{1+(DI\,\bar{T})} \tag{2.21}$$

ou, determinando a densidade de incidência a partir da prevalência,

$$DI = \frac{P}{1-P}\,\frac{1}{\bar{T}}, \tag{2.22}$$

lembrando que a suposição feita é de que a população em estudo é estável.

## 2.3 Medidas de efeito

As medidas estudadas na seção anterior visam quantificar a ocorrência da doença numa determinada população. No entanto, a maior preocupação da Epidemiologia está em estabelecer relações causais e quantificar a influência dos fatores detectados como sendo importantes na modificação do risco de adoecer. Para atingir este objetivo, a idéia básica é comparar a ocorrência de doença entre uma população exposta a 1 ou mais fatores e outra considerada de referência, isto é, não exposta ou exposta ao fator em níveis considerados "normais" ou basais.

Há basicamente 2 maneiras de se medir a influência do fator na ocorrência de doença: pela razão entre medidas de freqüência ou pela diferença entre elas. Alguns autores, como Rothman (1986 [19]) chamam as medidas de efeito calculadas pela diferença de medidas de efeito absoluto e as calculadas pela razão de medidas de efeito relativo.

### 2.3.1 Medidas de efeito por razão

A razão entre taxas médias é chamada de *razão de densidades de incidências* e é definida pela expressão:

$$RDI_i = \frac{DI_i}{DI_0} = \frac{d_i/L_i}{d_0/L_0}, \tag{2.23}$$

cujos elementos são identificados na Tabela 2.1 que mostra os dados obtidos pela observação de $k+1$ populações. O índice zero identifica a população de referência e os índices de 1 a $k$ as $k$ populações expostas a diferentes níveis do fator ou fatores.

A medida varia de 0 a infinito, sendo que valores menores que 1 indicam uma associação negativa entre exposição e doença ("proteção"), valores maiores que 1, uma associação positiva ("risco"). Valores em torno de 1 indicam ausência de associação, o que chamaremos de *hipótese de nulidade* ($H_0$).

A razão de medidas de risco é chamada de *risco relativo* e definida como

$$RR_i = \frac{IC_i}{IC_0} = \frac{d_i/n_i}{d_0/n_0} \tag{2.24}$$

Tabela 2.1: Notação usada na definição de medidas de efeito a partir de dados avaliados em termos de contagem de tempo-pessoa.

| | Categorias de exposição | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | ... | k | Total |
| Casos novos | $d_0$ | $d_1$ | ... | $d_k$ | D |
| Tempo-pessoa | $L_0$ | $L_1$ | ... | $L_k$ | L |

Tabela 2.2: Notação usada na definição de medidas de efeito a partir de dados obtidos por freqüências.

| | Categorias de exposição | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | ... | k | Total |
| Casos novos | $d_0$ | $d_1$ | ... | $d_k$ | D |
| Não casos | $b_0$ | $b_1$ | ... | $b_k$ | B |
| Total | $n_0$ | $n_1$ | ... | $n_k$ | N |

cujos elementos estão definidos na Tabela 2.2 que mostra a notação para dados obtidos a partir de freqüências.

O risco relativo também varia de 0 a $+\infty$ e o valor que indica não associação é a unidade. O risco relativo pode ser interpretado como quantas vezes é maior o risco de se adoecer devido à exposição em questão. Rothman (1986 [19]) apresenta uma variação desta medida que é $RR_i - 1$, que é chamada de *acréscimo de risco*. Esta medida nos informa em quanto aumenta o risco quando da exposição ao fator.

A comparação de prevalências se faz pela medida chamada *razão de prevalências* que é calculada de modo similar ao risco relativo.

### 2.3.2 Medidas de efeito por diferença

As medidas de efeito calculadas por diferença são encontradas pela subtração do valor encontrado para a população de referência do valor encontrado para o i-ésimo grupo exposto.

A *diferença de densidades de incidência* é definida como:

$$DDI_i = DI_i - DI_0 = \left(\frac{d_i}{L_i}\right) - \left(\frac{d_0}{L_0}\right) \tag{2.25}$$

com os elementos definidos na Tabela 2.1. O valor que corresponde à hipótese de efeito nulo é zero, sendo que a medida varia de $-\infty$ a $+\infty$.

A *diferença de incidências cumulativas*, ou *risco atribuível*, é definida como:

$$RA_i = IC_i - IC_0 = \left(\frac{d_i}{n_i}\right) - \left(\frac{d_0}{n_0}\right) . \tag{2.26}$$

Veja a Tabela 2.2.

Esta medida mostra qual a parcela do risco existente para o i-ésimo grupo é de responsabilidade exclusiva da exposição.

A *diferença de prevalências* é calculada de modo análogo ao risco atribuível.

### 2.3.3 A fração etiológica

Esta medida combina diferença e razão, de forma a dar uma medida do incremento absoluto na incidência da doença, proporcional à incidência

no i-ésimo grupo exposto. A fração etiológica é definida como

$$FE_i = \frac{DI_i - DI_0}{DI_i}. \tag{2.27}$$

## 2.4 Os estudos epidemiológicos

Pode-se desenvolver uma pesquisa usando basicamente dois tipos estudos, os experimentais e os não-experimentais ou observacionais. O primeiro tipo de estudo se caracteriza pelo fato do pesquisador ter o controle do processo, principalmente no que diz respeito à atribuição dos tratamentos às unidades experimentais. No segundo, ao contrário, o pesquisador não tem liberdade nesta atribuição, devendo se limitar a observar grupos que por vontade própria ou por contingência de uma situação se enquadram em um determinado tratamento. É muito simples realizar um experimento quando se quer estimar o efeito de um fertilizante no crescimento de uma planta. Não há qualquer problema em se escolher um pedaço de terra, delimitar parcelas, plantar aí a espécie escolhida e, aleatoriamente, determinar quais receberão o fertilizante e quais não. Na área biomédica, por outro lado, isto em geral não é simples. A não ser em alguns casos específicos, como pesquisa terapêutica, não é viável ou ético se atribuir tratamentos aos indivíduos do estudo. Não se pode escolher uma centena de pessoas para a seguir sortear metade delas, determinando-se que fumem 20 cigarros por dia nos próximos 10 anos. Provavelmente muitas delas se recusariam, mas há uma limitação ética evidente no caso. Não se pode deliberadamente expor pessoas a qualquer fator que sabida ou supostamente aumente seu risco de adoecer. Também não se pode realizar um experimento para verificar se a raça é um fator de risco para a hipertensão, por razões óbvias. Assim é que a maior parte da pesquisa biomédica é feita com base em estudos observacionais.

Há uma grande variedade de desenhos para estes estudos. Kleinbaum et alii (1982 [10]) relacionam 15 desses desenhos. Nos deteremos apenas nos 3 básicos, que são o *estudo prospectivo*, o *estudo transversal* e o *estudo retrospectivo*.

### 2.4.1 O estudo prospectivo

O estudo prospectivo — também chamado de estudo de coortes ou *follow-up* — é um desenho em que se conhece o estado de exposição aos fatores

em estudo de cada um dos indivíduos no início do período de observação. Estes indivíduos são então acompanhados por um determinado período de tempo, quando a ocorrência de um ou mais tipos de doenças será anotada. Na população a ser observada estarão apenas pessoas que reconhecidamente não apresentam o agravo de interesse, isto é, que podem vir a adoecer (ser *casos*) durante a observação (Kleinbaum et alii, 1982 [10]).

Um estudo prospectivo pode ser estritamente prospectivo, histórico ou misto. O primeiro caso é o prospectivo típico, onde as coortes (grupos de pessoas a serem acompanhadas) são montadas e observadas a partir deste momento, no presente. É o estudo que mais se assemelha a um experimento. No prospectivo histórico, todas as informações são obtidas através de registros, eventos que ocorreram no passado. Pode ocorrer também uma mistura, onde as coortes são formadas com informações pregressas e alguma observação ainda é feita nas pessoas remanescentes.

Existem duas qualidades importantes neste tipo de estudo. A primeira é a questão da temporalidade. Conhece-se a exposição de início e se observa em que momento ocorre a doença. Isto permite conhecer a evolução natural da doença. Pode-se saber não apenas quantas pessoas adoeceram durante os $n$ anos de observação, mas de que forma, isto é, se as mortes ocorreram rapidamente, no início do período, se gradualmente ou se muito lentamente, ao final da observação. O estabelecimento desta seqüência temporal é muito importante para a definição de uma etiologia. A outra qualidade reside no fato de que as densidades de incidência para cada nível de exposição, assim como as incidências cumulativas, são diretamente estimáveis (Rothman, 1986 [19]).

Os estudos prospectivos, por outro lado, são sabidamente ineficientes quando a doença em estudo é rara. Neste caso, para que seja observado um número mínimo de casos, o estudo deverá contar com grandes coortes ou se estender por um longo período de tempo. Este prolongamento do estudo acaba por causar outros problemas. O maior está no seguimento dos indivíduos, que tendem, com o passar do tempo, a sair do estudo devido a morte por outras causas, migração, mudança de endereço ou simplesmente recusa de se apresentar para exame. Outra questão é, para coortes fixas, o envelhecimento destas, o que provoca efeitos difíceis de serem estimados ou controlados.

### 2.4.2 O estudo transversal

Estudos transversais são executados a partir da escolha de apenas uma amostra da população alvo. Os indivíduos escolhidos são submetidos a inquérito, de forma a se conhecer sua condição em relação à doença e em relação aos fatores em estudo. Como toda a população é usada no estudo, incluindo pessoas doentes e saudáveis, e a informação obtida se refere a um momento (um corte) no tempo, a proporção de doentes obtida é a prevalência. Por isso este estudo também é chamado de estudo de prevalência. Uma característica importante do estudo transversal é que tanto o fator ou fatores em estudo como a situação em relação à doença de interesse são aleatórios. No processo de amostragem não se fixou nenhuma característica para dirigir o processo de amostragem, que foi feito de apenas uma população, diferente do que ocorre no prospectivo e no retrospectivo, onde pelo menos 2 subpopulações são amostradas. Esta particularidade confere algumas propriedades estatísticas aos estudos transversais, como a possibilidade de se medir *correlação* entre o fator e a doença. Segundo Kleinbaum et alii (1982 [10]) este tipo de estudo é freqüentemente utilizado para se estabelecer novas hipóteses etiológicas.

### 2.4.3 O estudo retrospectivo

O estudo retrospectivo, também chamado de estudo de caso-controle, parte da classificação dos indivíduos de acordo com sua situação em relação à doença em estudo. Os indivíduos são classificados como doentes (casos) ou não doentes (controles), podendo haver mais de um grupo de controles. A partir disto, a condição de exposição ao fator em estudo será determinada retrospectivamente.

Os estudos de caso-controle são, em geral, mais fáceis e mais baratos de serem realizados do que os prospectivos. Isso se deve, em grande parte, ao fato de que podem ser feitos com dados já coletados, registrados em arquivos como prontuários médicos ou atestados de óbito. Além disto, o período de estudo costuma ser menor e não há, de regra, observação ou seguimento de pacientes. A maior desvantagem deste tipo de investigação está na ausência de temporalidade, que dificulta o estabelecimento de uma relação causal. A estimação de medidas de freqüência também apresenta dificuldades. A primeira informação que se tem é em relação à condição do indivíduo em relação

à doença, ou seja, doente ou não doente. A partir disso, a informação sobre exposição é pesquisada. Portanto, a proporção estimável é a de expostos para as populações de casos e controles, ou seja a probabilidade de exposição, dada a condição de doença.

Sem nenhuma informação adicional é impossível estimar a probabilidade de adoecer condicional à exposição. Se a incidência cumulativa da doença na população geral é conhecida, se torna possível de estimar esta probabilidade usando propriedades da probabilidade condicional. Neste caso, em termos da estimação da incidência cumulativa, os estudos prospectivo e de caso-controle se equivalem.

Outra grande dificuldade do caso-controle é que casos e controles provèm de populações diferentes. Cabe, então, ao pesquisador assegurar a comparabilidade destes 2 grupos em relação a fatores de risco extrínsecos (variáveis de confundimento) e a outras fontes de distorção. Esta tarefa é complexa e muitas vezes polêmica. A idéia que norteia a escolha do grupo de referência é a que ele seja o mais próximo possível da população de casos, não havendo necessidade que este grupo seja representativo da população geral (Rothman, 1986 [19]).

## 2.5 Interação e confundimento

Estes dois conceitos são usados muito freqüentemente nos textos epidemiológicos, revestindo-se de grande importância. Não obstante, ainda são alvos de controvérsia, especialmente o de confundimento.

**Interação** é um conceito estatístico clássico e é usado em Epidemiologia em sua forma original. Suponha que 2 fatores, A e B, influenciem o aparecimento de determinada doença. Por simplicidade, admitamos que temos 2 níveis de cada fator, de forma que temos 4 grupos com diferentes níveis exposição: não exposto aos 2 fatores (grupo 00), exposto somente ao primeiro (grupo 10), exposto somente ao segundo (grupo 01) e duplamente exposto (grupo 11). Se, para cada nível do fator B, o efeito da exposição ao fator A for diferente, diz-se que há interação. O conceito de interação é fortemente dependente do modelo adotado (Kleinbaum et alii, 1982 [10]). Assim, dependendo de qual seja a medida de efeito escolhida, interação terá uma interpretação um pouco diferente.

Se a medida de efeito de interesse é o risco relativo, então ocorrerá interação quando os riscos relativos para o fator A em cada nível de B forem diferentes. Ou, ao contrário, se estes riscos relativos forem iguais, diz-se que não há interação. Se chamamos de $\theta_i$ a probabilidade de adoecer no grupo i, a hipótese de não interação será

$$\frac{\theta_{10}}{\theta_{00}} = \frac{\theta_{11}}{\theta_{01}} \tag{2.28}$$

que é equivalente à expressão

$$\frac{\theta_{11}}{\theta_{00}} = \frac{\theta_{10}}{\theta_{00}} \frac{\theta_{01}}{\theta_{00}} \tag{2.29}$$

que representa um modelo multiplicativo. Isto é, na ausência de interação, o risco relativo da dupla exposição é igual ao produto dos riscos relativos das exposições simples.

Se a medida de efeito escolhida é a diferença de riscos, interação é definida como riscos atribuíveis diferentes para o fator A em cada nível de B, o que leva à expressão

$$\theta_{10} - \theta_{00} = \theta_{11} - \theta_{01} \tag{2.30}$$

para definir a hipótese de não interação. Esta expressão é equivalente a

$$\theta_{11} - \theta_{00} = (\theta_{10} - \theta_{00}) + (\theta_{01} - \theta_{00}) \tag{2.31}$$

que corresponde a um modelo aditivo, ou seja, o risco atribuível à dupla exposição é igual à soma dos riscos atribuíveis às exposições simples na ausência de interação.

O conceito de **confundimento** ainda é alvo de alguma controvérsia. Uma definição simples, útil para um entendimento preliminar do conceito é dada por Kleinbaum et alii (1982 [10]). Ela diz que confundimento corresponde à ocorrência de tendência na estimativa da relação entre exposição e doença, que pode acontecer quando este efeito está misturado com o efeito de variáveis extrínsecas ao problema. Dependendo do tipo de associação entre as variáveis de confundimento e exposição e doença, a associação entre estas duas últimas pode ser super ou subestimada (Breslow & Day, 1980 [3]).

Um fator, para ser considerado como de confundimento, precisa preencher determinados requisitos. Primeiramente, o fator deve ser preditivo

da doença entre os não expostos, isto é, deve haver associação entre fator de confundimento e doença entre os não expostos. Além disso, deve haver associação entre o fator de confundimento e exposição, assim como entre o fator e a doença, condicionalmente ao controle de todos os outros fatores sendo estudados (Miettinen & Cook, 1981 [13]). Desta forma, uma variável de confundimento deve ser, por si só, preditora da doença, independentemente da exposição em estudo. Se, retirado o efeito da exposição, a associação entre fator de confundimento e doença desaparece, isto significa que esta associação depende inteiramente daquela entre exposição e doença, não caracterizando este fator como de confundimento.

# Capítulo 3

# Modelagem dos estudos epidemiológicos

Uma das principais preocupações em Epidemiologia é a determinação do risco de adoecer para um indivíduo em determinadas sub-populações e a comparação destes riscos, como já comentamos na seção 2.3, onde apresentamos as maneiras mais freqüentes de se comparar riscos.

Num estudo prospectivo o estado de exposição de cada indivíduo numa população a 1 ou mais fatores de interesse é conhecido. Tem-se também informação sobre outras variáveis que se quer controlar, como fatores de confundimento. O vetor que contém as informações sobre a condição de exposição e sobre estas outras variáveis de interesse, denominados de forma geral como fatores em estudo, para o i-ésimo indivíduo, será denominado $X_i$. Denominaremos por $Y_i$ a variável aleatória binária que indicará se o i-ésimo indivíduo adoece ou não durante a observação. Esta variável assume o valor 0 se o i-ésimo indivíduo não adoece e 1 se o i-ésimo indivíduo adoece. Por convenção, teremos $n$ indivíduos observados e $p$ fatores em estudo. Assim, podemos definir o risco de adoecer de um indivíduo como sendo

$$P(Y_i = 1 \mid X_i) = \theta_i. \tag{3.1}$$

O que se pretende é modelar a probabilidade $\theta_i$ de adoecer em função dos fatores conhecidos. A forma mais simples de se estabelecer esta relação é usando um modelo linear:

$$P(Y_i = 1 \mid X_i) = \theta_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \ldots + X_{ip}\beta_p \tag{3.2}$$

onde os $\beta$'s são os parâmetros do modelo e os $X_{ij}$ são valores conhecidos da exposição do i-ésimo indivíduo ao j-ésimo fator em estudo. Apesar de ser simples, este modelo apresenta algumas dificuldades. A mais séria (Cox, 1970 [5]) se refere ao fato de $\theta_i$ ser uma probabilidade e estar restrito ao intervalo $(0,1)$. É possível que se encontrem valores ajustados para $\theta_i$ fora deste intervalo, quando são calculados a partir dos estimadores de quadrados mínimos. Outro problema a ser considerado é que a variância dos $Y_i$, além de não ser constante, depende de $\theta_i$, que é a esperança de $Y_i$. Veja que $Y_i = 0$, 1 e $Y_i^2 = Y_i$ e portanto,

$$Var(Y_i) = E(Y_i^2) - E(Y_i)^2 = \theta_i(1-\theta_i). \tag{3.3}$$

Além disso, analisando a própria essência dos fenômenos em estudo, não é razoável supor que a probabilidade de adoecer varie de 0 a 1 linearmente numa determinada faixa de exposição. Uma variação de acordo com uma curva sigmóide, do tipo de uma curva dose-resposta, parece bem mais verossímil. De acordo com esta curva, a probabilidade de adoecer cresce lentamente com a exposição até um determinado valor, a partir do qual há um crescimento rápido, voltando a ser lento conforme a probabilidade se aproxima de 1. Observe a Figura 3.1.

## 3.1 O modelo logístico

O modelo que tem se mostrado mais adequado para a modelagem que se pretende é o Modelo Logístico. O nome vem da transformação que se faz da probabilidade de adoecer $\theta_i$:

$$logito(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right), \tag{3.4}$$

sendo, ao longo deste trabalho, log(.) o logaritmo natural.

O modelo logístico é linear no logito de $\theta_i$, sendo escrito como

$$\lambda_i = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \sum_{s=1}^{p} X_{is}\beta_s \tag{3.5}$$

para $i = 1, 2, \ldots, n$.

Vamos definir algumas matrizes de forma que nos seja possível utilizar uma notação mais simples.

$$\beta_{[(p+1)\times 1]} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \lambda_{(n\times 1)} = \begin{pmatrix} \log\left(\frac{\theta_1}{1-\theta_1}\right) \\ \vdots \\ \log\left(\frac{\theta_n}{1-\theta_n}\right) \end{pmatrix}$$

$$\boldsymbol{X}_{[n\times(p+1)]} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \tag{3.6}$$

Usando a notação matricial, o modelo logístico definido em (3.5) é escrito como

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \lambda_i = X_i'\boldsymbol{\beta}, \tag{3.7}$$

onde $X_i$ é um vetor que contém a i-ésima linha de $\boldsymbol{X}$.

Escrevendo a expressão acima em função de $\theta_i$,

$$\frac{\theta_i}{1-\theta_i} = \exp(X_i'\boldsymbol{\beta})$$

$$\theta_i = \exp(X_i\boldsymbol{\beta}) - \theta_i \exp(X_i'\boldsymbol{\beta})$$

$$\theta_i(1+\exp(X_i'\boldsymbol{\beta})) = \exp(X_i'\boldsymbol{\beta})$$

$$P(Y_i = 1 \mid X_i) = \theta_i = \frac{\exp(X_i'\boldsymbol{\beta})}{1+\exp(X_i'\boldsymbol{\beta})} \qquad (i = 1, 2, \dots, n). \tag{3.8}$$

## 3.2 A transformação logística

Para dar uma idéia melhor do que está acontecendo com os parâmetros neste modelo, faremos um rápido estudo da transformação logística, cujo gráfico pode ser visto na Figura 3.1. Para isto, usaremos sua forma geral

$$\lambda = \log\left(\frac{\theta}{1-\theta}\right), \qquad 0 < \theta < 1$$

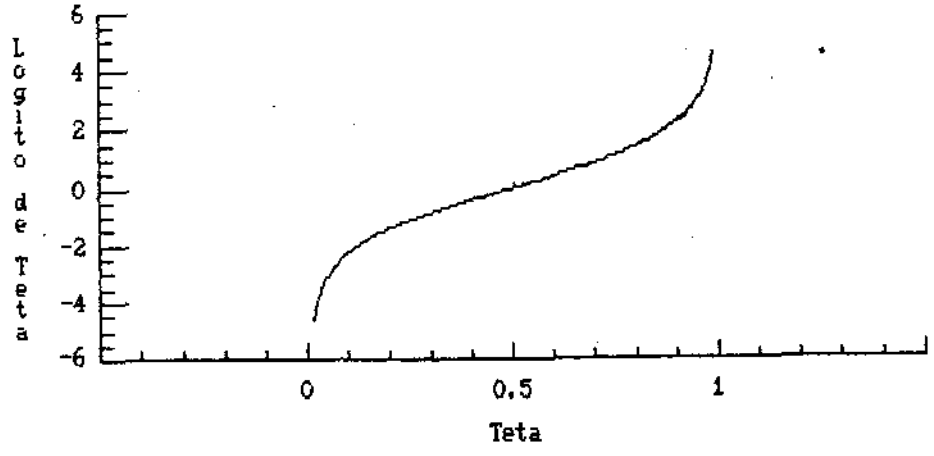


Figura 3.1: Gráfico do logito de $\theta$.

Esta função é também conhecida como o **logito** de $\theta$. Seu domínio é o intervalo $(0,1)$ e sua imagem é a reta real. Esta função é biunívoca, visto que sua derivada

$$\frac{d\lambda}{d\theta} = \frac{1}{\theta(1-\theta)}, \tag{3.9}$$

é sempre positiva para $\theta \in (0,1)$, o que caracteriza uma função estritamente crescente. Isto significa que cada ponto do intervalo $(0,1)$ é levado a apenas um ponto do intervalo $(+\infty, -\infty)$ pela função logito.

Pelo exposto, fica claro que esta transformação resolve o principal problema do modelo linear comum. O parâmetro $\lambda$ varia em todo o $\Re$, não apresentando mais os problemas de estimação que $\theta$ oferecia por estar restrito ao intervalo $(0,1)$.

## 3.3 Aplicação do modelo aos estudos epidemiológicos

O modelo logístico é muito versátil, permitindo que praticamente todas as situações de interesse do pesquisador possam ser modeladas. Neste

| | Exposição | |
|---|---|---|
| | sim | não |
| Doente | $d_1$ | $d_0$ |
| Não Doente | $n_1 - d_1$ | $n_0 - d_0$ |
| Total | $n_1$ | $n_0$ |

Tabela 3.1: Dados referentes a uma resposta e uma exposição binárias na forma de uma tabela $2 \times 2$.

trabalho nos restringiremos a estudar o modelo com uma variável resposta binária. Na literatura, alguns autores apresentam a extensão do modelo logístico para respostas politômicas (e.g. Prentice & Pyke, 1979 [17]). Mostraremos a seguir como pode ser construído este modelo para uma variedade de situações (a partir de Kleinbaum et alii, 1982 [9]), desde a mais simples, com apenas uma variável de exposição binária, até casos em que se incluem no modelo variáveis de estratificação, interação e confundimento. Inicialmente estaremos assumindo que temos um estudo prospectivo para analisar, onde os parâmetros de interesse, incluindo os riscos relativos, são diretamente estimáveis. Em seguida estenderemos estes resultados também aos estudos de caso-controle.

### 3.3.1 As diferentes parametrizações do modelo

**O caso $2 \times 2$**

Este é o caso mais simples, onde temos uma variável resposta binária, que indica se o i-ésimo indivíduo adoeceu ($Y_i = 1$) ou não ($Y_i = 0$) e apenas uma variável de exposição, também binária, que indica se o i-ésimo indivíduo faz parte do grupo exposto ($X_i = 1$) ou do não exposto ($X_i = 0$). Esta situação gera como resultado uma tabela de contingência $2 \times 2$, exemplificada na Tabela 3.1. Neste caso temos $n_1$ indivíduos expostos e $n_0$ indivíduos não expostos, $d_1$ indivíduos adoeceram entre os expostos e $d_0$ adoeceram entre os não expostos.

Assumimos que o evento de um indivíduo adoecer ou não durante o período de observação, representado pela variável aleatória $Y_i$, representa um ensaio de Bernoulli com probabilidade de sucesso $\theta_i$ (se estivéssemos consi-

derando um estudo de caso-controle a variável que teria caráter aleatório seria a que indica o nível de exposição). Supomos também que a probabilidade de adoecer é igual para os indivíduos dentro de cada um dos grupos de exposição. Como neste caso temos apenas 2 grupos, teremos apenas 2 probabilidades diferentes de adoecer:

$$\theta_{\bar{E}} = P(Y_i = 1 \mid X_i = 0) \tag{3.10}$$

e

$$\theta_E = P(Y_i = 1 \mid X_i = 1). \tag{3.11}$$

O modelo logístico neste caso é escrito da seguinte forma:

$$logito\,[P(Y_i = 1 \mid X_i)] = \beta_0 + \beta_1 X_i, \tag{3.12}$$

o que para $X_i = 0$ simplifica para a expressão

$$\lambda_{\bar{E}} = logito\,[P(Y_i = 1 \mid X_i = 0)] = logito(\theta_{\bar{E}}) = \beta_0 \tag{3.13}$$

e para $X_i = 1$,

$$\lambda_E = logito\,[P(Y_i = 1 \mid X_i = 1)] = logito(\theta_E) = \beta_0 + \beta_1 \tag{3.14}$$

Colocando as expressões acima em função dos parâmetros do modelo, temos que

$$\beta_0 = logito(\theta_{\bar{E}}) = \log\left(\frac{\theta_{\bar{E}}}{1-\theta_{\bar{E}}}\right) \tag{3.15}$$

e que

$$\beta_1 = logito(\theta_E) - logito(\theta_{\bar{E}})$$

$$\beta_1 = \log\left(\frac{\theta_E}{1-\theta_E}\right) - \log\left(\frac{\theta_{\bar{E}}}{1-\theta_{\bar{E}}}\right)$$

$$\beta_1 = \log\left(\frac{\theta_E(1-\theta_{\bar{E}})}{\theta_{\bar{E}}(1-\theta_E)}\right). \tag{3.16}$$

Isto é, o parâmetro $\beta_1$ corresponde ao log da *razão de odds* das probabilidades de adoecer entre os expostos e os não expostos. O parâmetro $\beta_0$ corresponde simplesmente ao logito da probabilidade de adoecer entre os não expostos.

Testar a ausência de associação entre exposição e doença significa testar que a probabilidade de adoecer entre os expostos é a mesma que aquela entre os não expostos, ou seja, $\theta_E = \theta_{\bar{E}}$, o que no modelo logístico corresponde a testar que $\beta_1 = 0$, que é equivalente a testar que a razão de odds em (3.16) é igual a 1.

Até agora vimos algumas medidas de efeito que são comparações entre densidades de incidência, incidências cumulativas ou prevalências. No modelo logístico surge uma outra medida, que é a razão de odds. Define-se como odds[1] de uma probabilidade o quociente entre esta probabilidade e seu complementar:

$$odds(p) = \frac{p}{1-p} \tag{3.17}$$

Chamaremos de razão de odds o quociente entre o odds da probabilidade de adoecer para um determinado grupo exposto e o odds da probabilidade de adoecer para o grupo de referência (não exposto). Assim, a razão de odds da exposição $E$ é

$$RO_E = \frac{\theta_E/(1-\theta_E)}{\theta_{\bar{E}}/(1-\theta_{\bar{E}})} = \frac{\theta_E(1-\theta_{\bar{E}})}{\theta_{\bar{E}}(1-\theta_E)}. \tag{3.18}$$

No caso de ausência de efeito da exposição, a razão de odds é igual a 1. Valores maiores que a unidade indicam risco e menores, proteção. Vários autores tentam comparar a razão de odds ao risco relativo (RR), tentando interpretá-la como uma aproximação deste. Mostra-se facilmente a partir de (3.18) que

$$RO_E = RR_E \times \frac{1-\theta_{\bar{E}}}{1-\theta_E}. \tag{3.19}$$

Portanto, quando se trabalha com doenças raras, a fração em (3.19) se aproxima de 1 e a razão de odds é uma boa aproximação do risco relativo. Independentemente de aproximar ou não o risco relativo, a razão de odds é uma medida de efeito que sempre concorda com ele em relação à direção do efeito. Ou seja, sempre que o risco relativo indicar um aumento do risco de adoecer ou uma diminuição deste risco, a razão de odds apontará na mesma direção, embora a amplitude do efeito indicada pelas 2 medidas possa ser diferente.

[1]Como ainda não se encontrou uma tradução apropriada em português, usaremos o termo original do inglês.

### Dois fatores de exposição binários

Neste caso estudamos dois fatores de exposição, $X_1$ e $X_2$, que podem assumir valores 1 ou 0, conforme o indivíduo esteja ou não exposto ao respectivo fator. Da mesma forma que no caso anterior, supomos que a probabilidade de adoecer em cada um dos grupos de exposição seja igual para todos os seus indivíduos. Definimos

$$\theta_{ij} = P(Y = 1 \mid X_1 = i, X_2 = j) \tag{3.20}$$

e

$$RO_{ij} = \frac{\theta_{ij}\,(1 - \theta_{00})}{\theta_{00}\,(1 - \theta_{ij})} \tag{3.21}$$

de forma que $\theta_{00}$ é a probabilidade de adoecer de um indivíduo não exposto aos dois fatores, portanto pertencente ao grupo considerado basal em termos da exposição.

O modelo logístico para este caso pode ser escrito como

$$\lambda_i = \beta_0 + X_{i1}\,\beta_1 + X_{i2}\,\beta_2. \tag{3.22}$$

No entanto, o maior interesse neste caso está em estudar a interação entre os dois fatores, que em (3.22) não foi considerada. Há várias formas de se modelar interação (definida na seção 2.5). A mais simples e que conduz a um modelo de interação multiplicativo de fácil interpretação é usada no modelo a seguir:

$$\lambda_i = \beta_0 + X_{i1}\,\beta_1 + X_{i2}\,\beta_2 + X_{i1}X_{i2}\,\beta_3. \tag{3.23}$$

Agora a pergunta que se faz é se a interação é diferente de zero, o que equivale a testar $\beta_3 = 0$. Calculando os parâmetros do modelo de forma similar ao caso anterior, encontramos que

$$\beta_0 = logito(\theta_{00}) \tag{3.24}$$

$$\beta_1 = \log(RO_{10}) \tag{3.25}$$

$$\beta_2 = \log(RO_{01}) \tag{3.26}$$

$$\beta_3 = \log\left(\frac{RO_{11}}{RO_{10}\,RO_{01}}\right) \tag{3.27}$$

de maneira que testar $\beta_3 = 0$ equivale a testar $RO_{11} = RO_{10} \times RO_{01}$. Esta expressão revela que o modelo logístico é um modelo multiplicativo, ou seja, os efeitos se combinam de forma multiplicativa na ausência de interação e a presença desta é sinal de um afastamento da situação de multiplicação dos efeitos.

### Um fator binário com estratificação

No planejamento e análise de estudos epidemiológicos é muito comum o uso de estratificação para controlar fatores que, embora não interessem ao pesquisador diretamente, podem influir no risco de adoecer e falsear conclusões se não forem devidamente controlados. Por exemplo, é muito comum que se estratifique por idade, um fator que muito freqüentemente altera a probabilidade de adoecer. Os fatores usados para estratificação não terão seus efeitos quantificados nem testados, visto que esta técnica serve para controlar ou retirar o efeito do fator. Assim é que os fatores usados para estratificação são aqueles cujos efeitos são conhecidos pelo pesquisador, que tem interesse de controlá-los seja para reduzir a variância do modelo, seja para eliminar os efeitos de confundimento do fator sobre outras variáveis em estudo.

Supomos que temos os indivíduos em estudo divididos em t estratos, com apenas um fator binário de risco. O modelo logístico, na forma mais geral, é escrito como

$$\lambda_{is} = \beta_{0s} + X_i \beta_{1s} \tag{3.28}$$

onde $i$ indica o indivíduo e $s$ o estrato. Este modelo propõe que não apenas o risco basal (do grupo não exposto) se altera de um estrato para outro ($\beta_{0s}$ diferentes), como também o efeito provocado pela exposição varia entre estratos ($\beta_{1s}$) diferentes. Isto é, a exposição tem um efeito de magnitude diferente dentro de cada estrato e a razão de odds entre expostos e não expostos para o estrato s será $\exp(\beta_{1s})$. Se supomos que o efeito da exposição é constante dentro dos estratos, havendo variação apenas no risco basal de adoecer, um modelo mais simples pode ser proposto:

$$\lambda_{is} = \beta_{0s} + X_i \beta_1. \tag{3.29}$$

Agora a razão de odds entre expostos e não expostos em cada extrato é constante, igual a $\exp(\beta_1)$, revelando que os riscos relativos se mantêm constantes entre estratos.

Cabe ressaltar que vários autores acreditam que a inclusão no modelo das variáveis usadas para estratificação resulta em análises mais eficientes, principalmente devido à redução do número parâmetros. Veja, por exemplo, que se tivermos 5 faixas etárias correspondendo a 5 estratos, no modelo mais simples teremos 6 parâmetros a serem estimados e no modelo completo, 10. A inclusão da idade no modelo como fator implicará no acréscimo de

4 parâmetros. Assim, o número de parâmetros será igual no primeiro caso, mas bem menor no segundo. Além disso, se a variável a ser controlada for contínua, ela pode ser incluída no modelo como tal, evitando a perda de informação que sempre ocorre quando se discretiza uma variável.

### Modelo com variáveis de confundimento e interações

Neste exemplo, incluiremos no modelo variáveis de confundimento e interações entre estas variáveis e a variável de exposição binária. As variáveis de confundimento podem ser contínuas ou discretas. Para maior clareza, usaremos a letra $E$ para a variável de exposição, que pode assumir os valores 0 ou 1 e $C_i$ para as $p$ variáveis de confundimento. As $q$ interações serão denotadas por $W_i$, funções de $E$ e das $C_i$. Por exemplo, $W_1 = E\,C_1$, $W_2 = E\,C_1\,C_2$. Retiramos também os índices relativos ao indivíduo, por simplicidade, insistindo que a relação deve ser considerada para cada um deles. O modelo é escrito como

$$\lambda = \beta_0 + E\,\beta + \sum_{i=1}^{p} C_i\,\gamma_i + \sum_{j=1}^{q} W_j\,\delta_j \tag{3.30}$$

e inclui $p + q + 2$ parâmetros.

Finalizando esta seção, observamos que embora nestes exemplos tenhamos usado apenas variáveis de exposição binárias, o modelo permite que sejam usadas variáveis discretas ou contínuas. Toda vez que uma variável discreta com $m$ níveis for utilizada, será necessário que se inclua no modelo $m-1$ parâmetros, sendo que a matriz de regressão, nas colunas correspondentes a estes parâmetros conterá 0 ou 1, de forma a incluir o parâmetro correto no modelo. O nível considerado basal será representado por zeros nas $m-1$ colunas da matriz. Por exemplo, uma exposição com 3 níveis necessitará de 2 parâmetros, $\beta_1$ e $\beta_2$. As respectivas colunas da matriz do modelo conterão $0,0$ para o nível basal, $1,0$ para o nível intermediário e $0,1$ para o nível alto da exposição. Se a variável for contínua, apenas um parâmetro será usado e a matriz do modelo conterá uma coluna com os valores observados da variável para cada indivíduo.

## 3.3.2 O modelo logístico em estudos de caso-controle

Nos estudos de caso-controle as probabilidades de adoecer não são diretamente estimáveis, como já mostramos anteriormente. Desta forma, não

é imediato supor que o modelo logístico, que propõe um modelo para uma função do risco de adoecer, seja aplicável a este tipo de estudo. Prentice e Pyke (1979 [17]) mostraram que não só o modelo logístico pode ser aplicado aos estudos de caso-controle, como, do ponto de vista deste modelo, eles são equivalentes aos estudos prospectivos.

Primeiramente, notemos que, embora os riscos não sejam estimáveis no caso-controle, existe uma medida de efeito que é: a razão de odds. No caso-controle, estima-se a probabilidade de se observar o vetor de exposição $\boldsymbol{X}$, dada a presença ou ausência de doença: $P(\boldsymbol{X} \mid Y)$. A razão de odds entre uma determinada exposição $\boldsymbol{X}$ e a exposição basal $\boldsymbol{X}_0$ é escrita como

$$RO_X = \frac{P(Y=1 \mid \boldsymbol{X})\,[1 - P(Y=1 \mid \boldsymbol{X}_0)]}{P(Y=1 \mid \boldsymbol{X}_0)\,[1 - P(Y=1 \mid \boldsymbol{X})]} \tag{3.31}$$

que, usando a relação da probabilidade condicional

$$P(Y=1 \mid X=x) = \frac{P(X=x \mid Y=1)\,P(Y=1)}{P(X=x)} \tag{3.32}$$

pode ser reescrita como

$$RO_X = \frac{P(\boldsymbol{X} \mid Y=1)\,P(\boldsymbol{X}_0 \mid Y=0)}{P(\boldsymbol{X} \mid Y=0)\,P(\boldsymbol{X}_0 \mid Y=1)} \tag{3.33}$$

sendo que todas estas probabilidades são estimáveis num estudo caso-controle. Lembrando que os parâmetros do modelo logístico são razões de odds ou funções delas, já se pode ter uma pista da aplicabilidade do modelo aos estudos de caso-controle.

Depois de mostrar esta equivalência, Prentice e Pyke (1979 [17]) deduzem o modelo induzido pelo modelo logístico para um estudo prospectivo quando os dados provêm de um estudo retrospectivo e mostram que este modelo também é da forma logística. Em seguida, calculam a equação de verossimilhança para os 2 modelos, de forma a poder compará-las. A verossimilhança obtida para o modelo com dados prospectivos é

$$L_p = \frac{\prod_{i=1}^{n_1} \exp(\beta_0 + X_i'\beta)}{\prod_{i=1}^{n} \left[1 + \exp(\beta_0 + X_i'\beta)\right]} \tag{3.34}$$

E para o modelo com dados retrospectivos temos

$$L_r = \frac{\prod_{i=1}^{n_1} \exp(\beta_0^* + X_i'\beta)}{\prod_{i=1}^{n} \left[1 + \exp(\beta_0^* + X_i'\beta)\right]} \prod_{i=1}^{n} P(X_i). \tag{3.35}$$

O resultado acima é mostrado nesta forma por Kleinbaum et alii (1982 [9]). Ele é fundamental, pois, daí se depreende que as verossimilhanças são proporcionais, isto é, a razão entre as 2 equações de verossimilhança é uma constante, não depende dos parâmetros. Isto significa que, essencialmente, a informação contida num estudo prospectivo e num caso-controle do ponto de vista do modelo logístico é a mesma. No processo de estimação por Máxima Verossimilhança a função de verossimilhança é maximizada em relação aos parâmetros, usando-se os dados da amostra obtida. Os valores encontrados neste processo serão usados como estimativa dos parâmetros de interesse. Assim, a função de verossimilhança é o único elo de ligação entre a amostra (espaço amostral) e os parâmetros (espaço paramétrico). Consideramos um experimento estatístico como a realização de um ensaio que inclui um determinado procedimento de amostragem. Se ocorre que 2 experimentos estatísticos diferentes apresentam funções de verossimilhança iguais ou proporcionais, isto significa que a informação que cada experimento traz sobre os parâmetros é equivalente à do outro. Ou seja, qualquer dos experimentos traz a mesma informação sobre os parâmetros de interesse. Prentice & Pyke (1979 [17]) mostram ainda que os estimadores dos parâmetros e da matriz de variância e covariância são os mesmos em ambos os modelos, como esperado.

Estes resultados nos permitem, de forma definitiva, usar o modelo logístico tanto para os estudos prospectivos quanto para os retrospectivos e ainda afirmam que estes estudos são equivalentes quando analisados por este modelo. Por essa razão, o modelo logístico é uma ferramenta muito poderosa na análise epidemiológica, reforçando a importância dos estudos de caso-controle e facilitando sua análise.

Por outro lado, se estes 2 tipos de estudos são equivalentes, pode-se questionar onde está a tão referida superioridade dos estudos prospectivos. Observemos que o modelo logístico não leva em conta a *temporalidade*, isto é, o momento de ocorrência dos eventos de interesse. Este é um dado que está disponível nos estudos prospectivos e é o que diferencia estes dos estudos de caso-controle. A informação de como os indivíduos adoecem no decorrer da observação não é levada em conta no modelo logístico, assim, se a maioria dos indivíduos adoece logo no início da observação ou só no fim do período, isto não altera o problema do ponto de vista deste modelo. Portanto, fica claro que é necessário um outro tipo de análise que leve em conta o tempo de observação até adoecer para que toda a informação contida num estudo prospectivo seja utilizada.

As 2 equações acima trazem uma pequena modificação da notação que estamos usando, separando o parâmetro $\beta_0$ da expressão $(X_i'\beta)$ para enfatizar que este parâmetro não é exatamente igual nos estudos prospectivo e caso-controle. Já vimos que, no prospectivo, $\beta_0$ é o logito da probabilidade de adoecer entre os não expostos. No caso-controle, este parâmetro está confundido com outras quantidades que dependem da fração amostral (Breslow & Day, 1980 [3]). Devido a este fenômeno, este parâmetro perde sua interpretabilidade neste caso.

# Capítulo 4

# Estimação, testes de hipótese e diagnóstico

## 4.1 A função de verossimilhança

### 4.1.1 Verossimilhança não condicional

Para encontrar os estimadores dos parâmetros do modelo logístico vamos estudar sua função de verossimilhança, buscando estatísticas suficientes. A função de verossimilhança é a função de probabilidade da distribuição vista como uma função dos parâmetros para respostas fixadas (Bickel & Doksum, 1977 [1]). A função de verossimilhança do estudo em questão é

$$L(\boldsymbol{\beta}, \mathbf{y}) = \prod_{\{i:Y_i=1\}} P(Y_i = 1 \mid X) \prod_{\{i:Y_i=0\}} P(Y_i = 0 \mid X) \qquad (4.1)$$

$$L(\boldsymbol{\beta}, \mathbf{y}) = \prod_{\{i:Y_i=1\}} \frac{e^{X_i'\boldsymbol{\beta}}}{1+e^{X_i'\boldsymbol{\beta}}} \prod_{\{i:Y_i=0\}} \frac{1}{1+e^{X_i'\boldsymbol{\beta}}} =$$

$$= \prod_{i=1}^{n} \left\{ \frac{e^{X_i'\boldsymbol{\beta}}}{1+e^{X_i'\boldsymbol{\beta}}} \right\}^{y_i} \prod_{i=1}^{n} \left\{ \frac{1}{1+e^{X_i'\boldsymbol{\beta}}} \right\}^{1-y_i} =$$

$$= \frac{\prod_{i=1}^{n} \left\{ e^{X_i'\boldsymbol{\beta}} \right\}^{y_i}}{\prod_{i=1}^{n} \left\{ 1+e^{X_i'\boldsymbol{\beta}} \right\}^{y_i} \prod_{i=1}^{n} \left\{ 1+e^{X_i'\boldsymbol{\beta}} \right\}^{1-y_i}} =$$

$$= \frac{\exp\{\sum_{i=1}^{n} y_i(X_i'\beta)\}}{\prod_{i=1}^{n}\left\{1+e^{X_i'\beta}\right\}}. \tag{4.2}$$

Se estabelecermos que os casos ($Y_i = 1$) são os primeiros $n_1$ indivíduos, podemos simplificar a equação anterior para

$$L(\beta, \mathbf{y}) = \frac{\prod_{i=1}^{n_1} e^{X_i'\beta}}{\prod_{i=1}^{n}\left\{1+e^{X_i'\beta}\right\}} \tag{4.3}$$

onde $\boldsymbol{X}$ e $\boldsymbol{\beta}$ são os definidos em (3.6) e $X_i$ é o vetor que contém a i-ésima linha de $\boldsymbol{X}$. Esta equação é a mesma daquela apresentada em (3.34) com pequenas alterações de notação.

Continuando a buscar as estatísticas suficientes retomamos a equação (4.2) e escrevemos

$$L(\beta, \mathbf{y}) = \frac{e^{(\boldsymbol{X}\beta)'\mathbf{Y}}}{\prod_{i=1}^{n}\left\{1+e^{X_i'\beta}\right\}} =$$

$$= \frac{e^{\beta'\boldsymbol{X}'\mathbf{Y}}}{\prod_{i=1}^{n}\left\{1+e^{X_i'\beta}\right\}} = \frac{e^{\beta'\mathbf{T}}}{\prod_{i=1}^{n}\left\{1+e^{X_i'\beta}\right\}} \tag{4.4}$$

onde

$$\mathbf{T}_{[(p+1)\times 1]} = \mathbf{X}'\mathbf{Y} = \begin{pmatrix} X_{(0)}'\mathbf{Y} \\ \vdots \\ X_{(p)}'\mathbf{Y} \end{pmatrix}, \tag{4.5}$$

$$T_s = X_{(s)}'\mathbf{Y} = \sum_{i=1}^{n} X_{is}Y_i \qquad (s = 0, 1, \ldots, p). \tag{4.6}$$

onde $X_{(j)}$ é o vetor que contém a j-ésima coluna de $\boldsymbol{X}$.

Pelo teorema da fatoração de Neyman (Bickel & Doksum, 1977 [1], pg. 65), as $T_s$ são estatísticas suficientes e a partir delas poderemos encontrar as estatísticas ótimas. Se comparamos o modelo logístico com o modelo de regressão linear, veremos que as estatísticas suficientes para este modelo envolvem estatísticas idênticas às $T_s$ e ainda a soma de quadrado do resíduo. Isso indica que o modelo logístico tem mais semelhanças com o modelo de regressão linear sob teoria normal do que poderia parecer à primeira vista (Cox, 1970 [5]).

Encontradas a função de verossimilhança e as estatísticas suficientes, encontrar os estimadores do vetor de parâmetros $\boldsymbol{\beta}$ é um problema de cálculo. A dificuldade apresentada pelo modelo logístico fica clara com a equação (4.4): ela não é linear nos parâmetros. Porisso não é possível uma solução explícita para os parâmetros, a não ser em casos especiais muito simples. Em geral, eles têm que ser encontrados através de métodos numéricos iterativos, como o método de Newton-Raphson, por exemplo. Não nos deteremos neste detalhe da estimação.

### Os estimadores de Máxima Verossimilhança na tabela $2 \times 2$

O modelo logístico no caso em que temos uma resposta binária e uma exposição binária, que resulta numa tabela $2 \times 2$, está descrito no Capítulo 3, a partir da página 29. Nesta seção mostraremos os estimadores de Máxima Verossimilhança dos parâmetros deduzidos nas equações (3.15) e (3.16).

As matrizes utilizadas para descrever o modelo são:

$$\boldsymbol{X}_{(n\times 2)} = \begin{pmatrix} 1 & 0 \\ \vdots & \\ 1 & 1 \\ \vdots & \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \boldsymbol{\lambda}_{(n\times 1)} = \begin{pmatrix} \lambda_{\hat{E}} \\ \vdots \\ \lambda_E \\ \vdots \end{pmatrix} \tag{4.7}$$

$$X_i' = (1 \quad 0), \qquad i = 1, 2, \ldots, n_0$$
$$X_i' = (1 \quad 1), \qquad i = n_0 + 1, \ldots, n$$

As estatísticas suficientes para este caso são:

$$T_s = X_{(s)}' \mathbf{Y} \qquad (s = 1, 2)$$

e portanto,

$$T_1 = \sum_{i=1}^{n} Y_i \qquad T_2 = \sum_{i=n_0+1}^{n} Y_i \tag{4.8}$$

Para encontrar os estimadores de Máxima Verossimilhança dos parâmetros, igualaremos as estatísticas suficientes às respectivas esperanças (Bickel & Doksum, 1977 [1], pg. 102 e 106).

$$T_1 = n_0 \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} + n_1 \frac{e^{\hat{\beta}_0 + \hat{\beta}_1}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1}} \tag{4.9}$$

$$T_2 = n_1 \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1}} \tag{4.10}$$

Agora note que $T_1 = d_0 + d_1$ (Tabela 3.1) e que $T_2 = d_1$. Daí podemos escrever

$$n_1 \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1}} = n_0 \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} + n_1 \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1}} - d_0$$

$$\frac{d_0}{n_0} = \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}}$$

$$e^{\hat{\beta}_0} = \frac{d_0}{n_0 - d_0} \tag{4.11}$$

$$\hat{\beta}_0 = \log\left(\frac{d_0}{n_0 - d_0}\right) \tag{4.12}$$

Substituindo em (4.10),

$$d_1 = n_1 \frac{e^{\hat{\beta}_1} \frac{d_0}{n_0-d_0}}{1+e^{\hat{\beta}_1} \frac{d_0}{n_0-d_0}}$$

$$e^{\hat{\beta}_1} = \frac{d_1\ (n_0 - d_0)}{d_0\ (n_1 - d_1)} \tag{4.13}$$

$$\hat{\beta}_1 = \log\left(\frac{d_1\ (n_0 - d_0)}{d_0\ (n_1 - d_1)}\right) \tag{4.14}$$

É interessante notar que os estimadores de Máxima Verossimilhança dos parâmetros descritos nas equações (3.15) e (3.16) são simplesmente a substituição dos estimadores de Máxima Verossimilhança de $\theta_i$ nas respectivas expressões.

### 4.1.2 Verossimilhança condicional

Uma alternativa à estimação dos parâmetros por Máxima Verossimilhança (não condicional) é a utilização da verossimilhança condicional. Esta opção é importante na medida em que a estimação apresentada na seção anterior se torna inadequada quando o número de parâmetros do modelo se torna muito grande em relação à quantidade de dados disponíveis. Tipicamente, isto acontece quando os dados são finamente estratificados, sendo

pequeno número de observações em cada estrato. Este tipo de estratificação é freqüente nos estudos de caso-controle, onde um grande número de fatores de confundimento têm de ser controlados. Para estes casos, a estimação através da verossimilhança condicional se torna indicada, em vista de que esta técnica permite que o parâmetro $\beta_0$ seja eliminado, diminuindo o número de parâmetros a serem estimados.

Consideramos um estudo de caso-controle, sendo $x_1, \ldots, x_n$ os vetores com os fatores em estudo observados para os $n = n_1 + n_0$ casos mais controles. A verossimilhança condicional é calculada a partir da probabilidade de que os primeiros $n_1$ dos vetores $x_i$ observados realmente sejam casos, dado que exatamente $n_1$ dos $n$ indivíduos observados são casos. Esta probabilidade pode ser escrita como

$$\frac{\prod_{i=1}^{n_1} P(Y_i = 1 \mid x_i) \ \prod_{i=n_1+1}^{n} P(Y_i = 0 \mid x_i)}{\sum_u \left\{ \prod_{i=1}^{n_1} P(Y_i = 1 \mid x_{ui}) \ \prod_{i=n_1+1}^{n} P(Y_i = 0 \mid x_{ui}) \right\}} \tag{4.15}$$

onde o numerador é a probabilidade de se observar esta seqüência específica de casos e controles e exatamente $n_1$ casos. Ou seja, o numerador de uma probabilidade condicional ($P[A \cap B]$). Como o evento "observar exatamente esta seqüência de $n_1$ casos e $n_0$ controles" implica (está contido) no evento "obter exatamente $n_1$ casos", a probabilidade em questão é a mesma daquele de se observar a referida seqüência. O denominador é a probabilidade de se obter uma seqüência com exatamente $n_1$ casos, que é calculada somando-se as probabilidades da ocorrência de todas as seqüências que satisfazem esta condição. A soma em $u$ da equação tem este significado, isto é, a soma para todas as $\binom{n}{n_1}$ seqüências que contém os $n_1$ casos. Usando a transformação logística (3.8) obtemos a forma final da verossimilhança condicional,

$$L_c(\boldsymbol{\beta}, \mathbf{y}) = \frac{\prod_{i=1}^{n_1} \exp(X_i^{*\prime} \boldsymbol{\beta}^*)}{\sum_u \left\{ \prod_{i=1}^{n_1} \exp(X_{ui}^{*\prime} \boldsymbol{\beta}^*) \right\}} \tag{4.16}$$

onde $X_i^*$ é a i-ésima linha da matriz $\boldsymbol{X}$ sem o primeiro 1 e $\boldsymbol{\beta}^*$ o vetor $\boldsymbol{\beta}$ sem $\beta_0$.

Notemos que a verossimilhança condicional não depende de $\beta_0$, que foi eliminado pelo processo de permutação. Assim, ela permite a eliminação de um parâmetro molesto (*nuisance*), no estudo caso-controle, e maior acurácia

na estimação dos parâmetros de interesse. Por outro lado, o cálculo do denominador fica rapidamente inviabilizado conforme crescem os valores de $n$ e $n_1$. Mas nesse caso, com amostras de tamanho considerável, o processo anterior de estimação é confiável. Desta forma a estimação dos parâmetros do modelo logístico parece estar resolvida, dependendo agora apenas de algoritmos eficientes que permitam a resolução numérica do problema.

## 4.2 Qualidade de ajuste e escolha do modelo

Uma vez obtidos os estimadores de Máxima Verossimilhança, o passo seguinte é usá-los para fazer inferências estatísticas sobre as relações entre exposição e doença em estudo. Nesse momento, há ainda 2 tarefas que o pesquisador deve realizar. Em primeiro lugar, deve-se escolher um determinado modelo entre todos os possíveis. Em geral, parte-se do modelo completo, ou saturado, que contém todos os efeitos principais e interações e tenta-se simplificá-lo ao máximo, eliminando os parâmetros responsáveis pelas interações e efeitos que não se mostrarem importantes. Pode-se também partir de um modelo mínimo e ir-se adicionando fatores até que os restantes não sejam relevantes, ou mesmo usar uma técnica que misture estes dois procedimentos. Qualquer que seja o método escolhido, a maneira de se decidir se um parâmetro fica ou não no modelo final é semelhante. Para tal, é necessário que façam testes de hipótese, em geral seqüenciais, sobre os parâmetros. Existem vários procedimentos para realizar estes testes. Discutiremos a seguir a estatística de Wald, a estatística dos escores eficientes e a estatística da razão de verossimilhança.

Suponhamos que o vetor de $p+1$ parâmetros do modelo está particionado em dois sub-vetores com $q$ e $r$ parâmetros respectivamente, de forma que $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ e que temos interesse em testar a hipótese $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. Denotaremos por $\hat{\boldsymbol{\beta}}$ o vetor com os estimadores de Máxima Verossimilhança dos parâmetros do modelo sem restrições e $\hat{\boldsymbol{\beta}}_{H_0}$ o vetor de estimadores de Máxima Verossimilhança dos parâmetros do modelo sob $H_0$.

O vetor dos *escores eficientes* é definido como o vetor que contém as derivadas primeiras da função de verossimilhança em relação aos parâmetros,

$$U(\boldsymbol{\beta}) = \frac{\delta}{\delta \boldsymbol{\beta}} \log L(\boldsymbol{\beta}) \tag{4.17}$$

onde $L(\boldsymbol{\beta})$ é a função de verossimilhança para um modelo com vetor de parâmetros $\boldsymbol{\beta}$.

A *matriz de informação de Fisher* é definida como

$$I(\boldsymbol{\beta}) = -E\left(\frac{\delta}{\delta\boldsymbol{\beta}}U(\boldsymbol{\beta})\right) = V^{-1}(\hat{\boldsymbol{\beta}}), \tag{4.18}$$

onde $V^{-1}(\hat{\boldsymbol{\beta}})$ é a matriz de variância e covariância de $\hat{\boldsymbol{\beta}}$. Denotaremos por $I(\hat{\boldsymbol{\beta}})$ a matriz de informação observada, isto é, quando $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

Em seguida apresentamos as estatísticas que são usadas para testar a hipótese $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$.

### A estatística de Wald

A estatística de Wald é definida como

$$\hat{\boldsymbol{\beta}}_2\, \hat{V}_{22}^{-1}(\hat{\boldsymbol{\beta}})\, \hat{\boldsymbol{\beta}}_2 \tag{4.19}$$

onde $\hat{V}_{22}^{-1}(\hat{\boldsymbol{\beta}})$ é a porção da matriz de variância e covariância correspondente aos $r$ parâmetros que estão sendo testados. Esta estatística tem uma distribuição assintótica $\chi^2$ com $r$ graus de liberdade sob $H_0$.

### A estatística dos escores eficientes

Esta estatística é definida como

$$U'(\hat{\boldsymbol{\beta}}_{H_0})\, I^{-1}(\hat{\boldsymbol{\beta}}_{H_0})\, U(\hat{\boldsymbol{\beta}}_{H_0}) \tag{4.20}$$

que também tem distribuição assintótica $\chi^2$ com $r$ graus de liberdade sob $H_0$.

### A estatística da razão de verossimilhança

Esta estatística é a mais utilizada das 3, sendo definida como

$$G^2_{(\boldsymbol{\beta}_{H_0})} = -2\log\left(\frac{L(\hat{\boldsymbol{\beta}}_{H_0})}{L(\hat{\boldsymbol{\beta}})}\right). \tag{4.21}$$

Esta estatística também tem distribuição assintótica $\chi^2$ com $r$ graus de liberdade sob $H_0$. Ela é apresentada por Pregibon (1981 [16]) com o nome de desvio (*deviance*, D).

Quando se ajustam modelos em seqüência, especificamente modelos hierárquicos embutidos, as diferenças entre as estatísticas $G^2$ dos modelos também têm distribuição $\chi^2$, de forma que estas diferenças podem ser usadas para a escolha do melhor modelo para um determinado problema.

Modelos hierárquicos são aqueles que contêm todos os termos de menor ordem relacionados com um de ordem superior presente no modelo. Por exemplo, se a interação $X_1 X_2$ está presente, obrigatoriamente os fatores $X_1$ e $X_2$ estarão. Ou, se a interação $X_1 X_2 X_3$ estiver presente, também estarão as interações $X_1 X_2$, $X_1 X_3$, $X_2 X_3$ e os respectivos efeitos principais. Estes modelos são quase sempre utilizados na análise de dados por questões de interpretação, pois, se as interações forem analisadas em termos de contrastes, elas envolverão sempre contrastes dos efeitos principais (Fienberg, 1977 [6]).

Modelos hierárquicos embutidos são conjuntos de modelos contidos um no outro, seqüencialmente. Por exemplo,

$$\begin{array}{ll} (1) & \lambda = \beta_0 + X_1\,\beta_1 + X_2\,\beta_2 + X_1\,X_2\,\beta_3 \\ (2) & \lambda = \beta_0 + X_1\,\beta_1 + X_2\,\beta_2 \\ (3) & \lambda = \beta_0 + X_1\,\beta_1 \end{array} \tag{4.22}$$

são modelos hierárquicos embutidos. Neste caso, a diferença $G^2_{(3)} - G^2_{(2)}$ tem distribuição $\chi^2$ com 1 grau de liberdade. E $G^2_{(3)} - G^2_{(1)}$ tem distribuição $\chi^2$ com 2 graus de liberdade.

Agora observemos que a estatística $G^2$ do modelo (3) é escrita como

$$G^2_{(3)} = -2\log\left(\frac{L\,(3)}{L\,(1)}\right) \tag{4.23}$$

e que a estatística $G^2$ de (2) é escrita como

$$G^2_{(2)} = -2\log\left(\frac{L\,(2)}{L\,(1)}\right) \tag{4.24}$$

e que a diferença entre estas estatísticas é

$$G^2_{(3)} - G^2_{(2)} = -2\log\left(\frac{L\,(3)}{L\,(1)}\right) - \left[-2\log\left(\frac{L\,(2)}{L\,(1)}\right)\right] \tag{4.25}$$

ou seja,

$$-2\log L\,(3) - (-2\log L\,(2))\,. \tag{4.26}$$

Isto significa que a diferença entre as estatísticas da razão de verossimilhança para 2 modelos embutidos é igual à diferença entre seus logaritmos da verossimilhança do modelo. O log da verossimilhança não tem uma distribuição determinada, mas, as diferenças, como já vimos, têm.

## 4.3 Diagnóstico no modelo logístico

Nesta seção abordamos de forma objetiva as idéias básicas que nos permitem fazer diagnóstico em regressão logística. Apresentaremos basicamente técnicas e medidas sugeridas por Pregibon (1981 [16]). Artigos mais recentes podem ser consultados para um aprofundamento no assunto: Landwehr, Pregibon & Shoemaker (1984 [11]), Kay & Little (1986 [8]) e Williams (1987 [20]).

Após se ter ajustado um modelo logístico a um conjunto de dados, o interesse que se tem é de avaliar se este modelo está adequado. Basicamente, podem ocorrer dois problemas: podem existir observações que são muito mal ajustadas, isto é, o valor observado está distante do ajustado, ou podem existir observações que, sozinhas, influenciam o ajuste de forma radical. Para o primeiro caso, denominamos estas observações de *aberrantes* e observamos que o problema se localiza na resposta, em geral se constituindo de uma resposta distante das outras. Para o segundo caso, chamamos as observações de *influentes*. O problema, agora, se localiza no espaço de desenho (matriz $X$) e em geral se constitui de um vetor de fatores distante dos demais, de forma que provoca fortes alterações no ajuste.

Nos modelos lineares sob teoria normal, já se tem bem estabelecido um conjunto de procedimentos para diagnóstico e se conhece bem a dinâmica dos problemas. Para se detectar pontos aberrantes usam-se, basicamente, os resíduos, que neste caso têm definição clara. Grandes resíduos indicam observações aberrantes, que, em geral, não têm grande influência sobre o ajuste. Os pontos influentes têm a característica de trazerem os valores ajustados para sua vizinhança, de forma que, geralmente, apresentam pequenos resíduos, mas, sua representação na matriz de projeção é importante. Veremos a seguir que o comportamento do modelo logístico é semelhante aos de teoria normal nestes aspectos.

O objetivo desta abordagem é apresentar medidas que permitam a identificação destas observações problemáticas a partir de quantidades que

sejam disponíveis num ajuste por Máxima Verossimilhança através dos pacotes estatísticos mais comuns. Desta forma, mesmo que as medidas não sejam calculadas diretamente pelo programa, pode-se determiná-las com alguma manipulação dos elementos disponíveis.

### 4.3.1 Resíduos e matriz de projeção em regressão logística

Para o diagnóstico em regressão logística, parte-se de elementos equivalentes aos resíduos e à matriz de projeção dos modelos sob teoria normal. No nosso caso, os resíduos não são definidos de forma única. As duas formas mais interessantes, segundo Pregibon (1981 [16]), são os componentes do $\chi^2$ de Wald e os componentes da estatística da razão de verossimilhança. Para a notação supomos que temos $k+1$ populações (ou caselas) que contêm $n_i$ $(i = 0, 1, \ldots, k)$ elementos, de forma que $\sum_{i=0}^{k} n_i = n$ e que a resposta $Y$ é binomial, podendo variar de 0 a $n_i$ (veja a Tabela 2.2).

O $\chi^2$ de Wald pode ser escrito como

$$\chi^2 = \sum_{i=0}^{k} \frac{(d_i - n_i \hat{\theta}_i)^2}{n_i \hat{\theta}_i (1 - \hat{\theta}_i)} \tag{4.27}$$

sendo que seus componentes são

$$\chi_i = \frac{d_i - n_i \hat{\theta}_i}{\sqrt{n_i \hat{\theta}_i (1 - \hat{\theta}_i)}} \tag{4.28}$$

para $i = 0, 1, \ldots, k$. Segundo o modelo logístico $\hat{\theta}_i = \exp(X_i' \hat{\beta})/[1+\exp(X_i' \hat{\beta})]$.

A estatística da razão de verossimilhança pode ser escrita na forma

$$G^2 = \sum_{i=0}^{k} g_i^2 = \sum_{i=0}^{k} -2 \left[\log\, L_i(\hat{\theta}_i) - \log\, L_i(\hat{\theta}_i^s)\right] \tag{4.29}$$

onde $\hat{\theta}_i^s$ são os parâmetros do modelo irrestrito (ou saturado). Para cada população, $\hat{\theta}_i^s = d_i/n_i$. O log da verossimilhança pode ser calculado pela expressão

$$\log\, L_i(\hat{\theta}) = d_i \log \hat{\theta} + (n_i - d_i) \log(1 - \hat{\theta}). \tag{4.30}$$

Os componentes $g_i$ são calculados como

$$g_i = (-1)^{(\hat{\theta}_i^s < \hat{\theta}_i)} \sqrt{2} \sqrt{\log\, L_i(\hat{\theta}_i) - \log\, L_i(\hat{\theta}_i^s)} \tag{4.31}$$

para $i = 0, 1, \ldots, k$. O sinal do componente é determinado pelo fato de o valor observado da probabilidade da casela ser maior $(+)$ ou menor $(-)$ do que o valor ajustado. Isto é determinado na expressão acima através da desigualdade lógica do expoente de $-1$, que assume o valor 1 quando verdadeira e 0 quando falsa.

A matriz de projeção análoga para o modelo logístico é

$$\mathbf{H} = \mathbf{V}^{1/2}\mathbf{X}\,(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{V}^{1/2} \tag{4.32}$$

que é exatamente igual à matriz que aparece num ajuste por quadrados mínimos ponderados e onde $\mathbf{V}$ é uma matriz diagonal que contém as variâncias dos $y_i$. Ou seja, $v_{ii} = n_i\,\hat{\theta}_i(1-\hat{\theta}_i)$. A diferença importante que há entre este modelo e o de quadrados mínimos ponderados é que aqui a matriz $\mathbf{V}$ é obtida a partir do ajuste e não tem qualquer influência nele, enquanto que lá, esta matriz dá pesos às observações, diminuindo a influência daquelas que tenham grande variância.

Do ponto de vista computacional é interessante que se faça uma transformação na expressão anterior. Vamos fazer $\mathbf{W} = \mathbf{V}^{1/2}\mathbf{X}$. Reescrevendo a expressão, temos

$$\mathbf{H} = \mathbf{W}\,(\mathbf{W}'\mathbf{W})^{-1}\,\mathbf{W}' \tag{4.33}$$

de forma que podemos encontrar os elementos da diagonal da matriz de projeção usando a seguinte expressão:

$$h_{ii} = W_i\,(\mathbf{W}'\mathbf{W})^{-1}\,W'_{(i)} \tag{4.34}$$

onde $W_i$ é a i-ésima linha de $\mathbf{W}$ e $W'_{(i)}$ é a i-ésima coluna de $\mathbf{W}'$. O cálculo direto dos elementos da diagonal de $\mathbf{H}$ reduz muito o uso de memória do computador e também o tempo de processamento, visto que se formos calcular a matriz de projeção inteira, estaremos trabalhando com uma matriz quadrada de dimensão $k+1$.

De posse dos resíduos e da matriz de projeção, podemos agora fazer uma análise semelhante ao caso de teoria normal. Grandes resíduos devem indicar pontos mal ajustados ou aberrantes. Da mesma forma, grandes valores da diagonal de $\mathbf{H}$ devem indicar observações influentes no ajuste. Segundo Pregibon (1981 [16]), estas regras são confirmadas na prática, podendo ser utilizadas para o diagnóstico de observações influentes ou aberrantes.

### 4.3.2 Perturbações no modelo

Embora as medidas apresentadas na seção anterior nos dêem informações importantes, elas são incapazes de mostrar com clareza o impacto de cada observação nos diversos aspectos do modelo ajustado. Desta forma, desenvolvemos a seguir a idéia de causar pequenas perturbações no modelo escolhido, de forma a avaliar com maior detalhe a importância de cada observação (população ou casela).

A verossimilhança do modelo pode ser reescrita de forma a permitir que sejam dados pesos a cada observação:

$$L_w(\beta, \mathbf{y}) = \prod_{i=0}^{k} w_i \, L(\beta, y_i). \tag{4.35}$$

Assim, é possível introduzir pequenas alterações no modelo variando-se o valor dos $w_i$. Para se estudar uma determinada observação, fazemos

$$w_i = \begin{cases} w & para\ i = m \\ 1 & caso\ contrário \end{cases} \tag{4.36}$$

com $w$ variando entre 0 e 1.

Os novos parâmetros do modelo são obtidos maximizando a verossimilhança (4.35) através do método usual. O problema desta abordagem é que o custo computacional é elevado, principalmente quando o número de observações é relativamente grande, visto que para estudar todas as $k+1$ observações teremos que ajustar $k+1$ modelos. Na tentativa de reduzir este custo, propõe-se que sejam usadas estimativas dos parâmetros obtidas depois de apenas uma iteração do processo, que é iniciado com os estimadores do modelo usual. A grande vantagem destes estimadores, chamados estimadores de um passo, é que eles podem ser determinados a partir de expressões fechadas de dados disponíveis do ajuste inicial. O vetor de estimadores de um passo obtidos da perturbação da m-ésima observação é determinado através da expressão

$$\hat{\beta}^1(w) = \hat{\beta} - \frac{(X'\mathbf{V}\mathbf{W})^{-1} X_m s_m (1-w)}{[1-(1-w)h_{mm}]} \tag{4.37}$$

onde W é uma matriz diagonal que contém os $w_i$ como descritos em (4.36) e $s_m = y_m - n_m \hat{\theta}_m$.

Se queremos estudar o efeito da retirada de uma observação no ajuste, igualamos $w_m$ a 0 e a variação no vetor de parâmetros provocada por esta retirada é

$$\Delta_m \hat{\boldsymbol{\beta}}^1 = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1} X_m s_m/(1 - h_{mm}). \tag{4.38}$$

Pregibon (1981 [16]) propõe a avaliação do gráfico de $\Delta_m \hat{\beta}_j^1/ep(\hat{\beta}_j) \times m$. Através deste gráfico se consegue identificar as observações que causam instabilidades nos parâmetros $\hat{\beta}_j$.

Embora muito tempo do computador tenha sido economizado usando-se estes estimadores de um passo, observe que se o número de parâmetro do modelo é grande, teremos também muitos gráficos para examinar, visto que temos um para cada parâmetro. Desta forma, seria útil encontrar uma medida que informasse sobre a influência de cada observação no conjunto dos parâmetros. De posse desta medida, apenas um gráfico daria idéia de quais observações causam instabilidades no vetor de parâmetros.

A medida proposta se baseia na idéia de que

$$-2[\log L(\boldsymbol{\beta}, \mathbf{y}) - \log L(\hat{\boldsymbol{\beta}}, \mathbf{y})] = c \tag{4.39}$$

descreve o limite de uma região de confiança para o parâmetro $\boldsymbol{\beta}$. O cálculo desta equação no ponto $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(0)$ produz a medida $c_m$, da influência do ponto $m$ no vetor de parâmetros $\boldsymbol{\beta}$.

Realizando as aproximações necessárias e usando $\hat{\beta}^1(0)$ como estimador de $\boldsymbol{\beta}(0)$, apresentamos a medida

$$c_m^1 = \frac{\chi_m^2 \, h_{mm}}{(1 - h_{mm})^2}. \tag{4.40}$$

Novamente se sugere a análise da medida em gráfico de $c_m^1 \times m$ ou ainda um gráfico contra os valores ajustados $(\hat{\lambda})$.

As medidas apresentadas aqui podem ser calculadas sem grande dificuldade a partir da matriz de desenho $\boldsymbol{X}$, do vetor de respostas $\mathbf{Y}$ e do vetor de estimadores de Máxima Verossimilhança dos parâmetros $\hat{\boldsymbol{\beta}}$. Um programa que permita a execução de operações matriciais, como o IML (*Interactive Matrix Language*) do SAS, permite que os cálculos sejam feitos usando praticamente as mesmas expressões mostradas aqui. No capítulo 5 apresentaremos algumas sugestões que possam facilitar ao usuário não especializado o cálculo de tais medidas, de forma a incentivar que os modelos ajustados passem por uma análise crítica mais severa.

# Capítulo 5

# Modelo logístico – uma aplicação

Neste capítulo apresentamos a análise de um conjunto de dados real com o intuito de ilustrar as técnicas apresentadas e apresentar um roteiro básico de análise pelo qual pesquisadores não estatísticos possam se guiar ao realizar modelagens em seus dados. O conjunto escolhido contém dados de uma fase de uma grande pesquisa sobre AIDS realizada na Califórnia, EUA. Os dados foram gentilmente cedidos pelos pesquisadores da Universidade da Califórnia para utilização didática. O estudo que gerou os dados é um estudo retrospectivo. Os casos são os pacientes que apresentaram sorologia positiva para o vírus da AIDS e os controles aqueles com sorologia negativa. Como exposição foram estudadas originalmente 9 variáveis que descrevem o uso de agulhas compartilhadas para injeção endovenosa, exposição a transfusões de derivados de sangue, número de parceiros sexuais e várias práticas sexuais dos entrevistados. Para tornar a análise um pouco mais simples, algumas variáveis que se mostraram menos importantes em uma avaliação inicial foram retiradas e ficamos com 6 variáveis de exposição.

Toda a análise foi realizada através do pacote SAS, versão 6.03, instalado em um microcomputador PC–XT. Também foi utilizado, em alguns momentos, o SAS instalado em um computador de grande porte, IBM 3090, através de ligação micro–grande porte gerenciada pelo próprio SAS do micro. Vale ressaltar que o uso do grande porte se deu apenas com o intuito de reduzir o tempo de processamento, visto que todos os programas rodados são suportados pela versão do PC. Os modelos foram ajustados usando-se o

procedimento CATMOD. Para o cálculo da diagonal da matriz de projeção, usada para diagnóstico, foi usado o módulo IML que permite que operações matriciais sejam executadas diretamente, de forma interativa.

## 5.1 Descrição do conjunto de dados

O arquivo contém informação de 816 homens entrevistados. Destas, 777 observações são utilizáveis, visto que não há falta de informação em nenhuma das variáveis. O agravo em estudo é o resultado da sorologia para AIDS (HIV), que foi anotada como negativo (0) ou positivo (1). Dos 6 fatores de risco estudados, 5 são binários e um, número de parceiros sexuais nos últimos 2 anos, quantitativo. Os nomes e a descrição destas variáveis se encontra na Tabela 5.1. As variáveis binárias foram codificadas sempre como 0 para a resposta negativa e 1 para a afirmativa.

Originalmente as variáveis que aqui apresentamos como binárias estavam subdivididas em 5 classes. Por exemplo, a variável que informa sobre o entrevistado ter sido penetrado pelo seus parceiros admitia as respostas nenhum, um, alguns, quase todos ou todos. Mas, desta forma, a resposta fica relativizada, pois o indivíduo pode responder que foi penetrado por todos os parceiros e o número de parceiros é 1, enquanto que outro que dá a mesma resposta teve muitos parceiros. Então preferimos adotar a versão binária destas variáveis, de forma que elas indicam uma prática e a questão quantitativa fica por conta da variável que informa o número de parceiros.

## 5.2 Ajuste e escolha do modelo

Inicialmente ajustamos um modelo contendo todos os efeitos principais das variáveis em estudo. O resultado é claramente insatisfatório, o que se conclui pela estatística $G^2$, de valor muito elevado (modelo 1 da Tabela 5.2). A primeira tentativa de melhorar o ajuste foi no sentido de fazer uma transformação na variável número de parceiros (PARCEIRO). Estas contagens, em geral, têm distribuição muito assimétrica, concentrada em torno dos valores mais baixos. Assim, tentamos uma transformação logarítmica. A variável PARCLOG foi calculada como log(PARCEIRO+1). A adição de 1 ao número de parceiros é feita para se evitar problemas numéricos ao calcular, eventualmente, o logaritmo de 0. A avaliação do efeito da transformação pode ser

Tabela 5.1: Descrição das variáveis de exposição utilizadas no exemplo.

| Nome da variável | Descrição | Tipo |
|---|---|---|
| PARCEIRO | Número de parceiros nos últimos 2 anos | quantitativa |
| AGULHA | Agulha para injeção endovenosa compartilhada nos últimos 5 anos | binária |
| RETOENTB | Parceiro penetrou no reto do entrevistado nos últimos 2 anos | binária |
| RETOPARB | Entrevistado penetrou no reto do parceiro nos últimos 2 anos | binária |
| MAORETB | Introdução da mão ou punho no reto do entrevistado nos últimos 2 anos | binária |
| ARTRETB | Introdução de artefatos no reto do entrevistado nos últimos 2 anos | binária |

Tabela 5.2: Valores da estatística da razão de verossimilhança para os modelos ajustados.

| Modelo | Variáveis presentes | GL | $G^2$ | PROB |
|---|---|---|---|---|
| 1 | PARCEIRO, AGULHA, RETOENTB, RETOPARB, MAORETB, ARTRETB | 244 | 306.31 | 0.0041 |
| 2 | PARCLOG, AGULHA, RETOENTB, RETOPARB, MAORETB, ARTRETB | 244 | 263.57 | 0.1859 |
| 3 | PARCLOG, AGULHA, RETOENTB, RETOPARB, MAORETB, ARTRETB, MAOARTB | 243 | 260.54 | 0.2098 |
| 4 | PARCLOG, AGULHA, RETOENTB, RETOPARB, MAOARTB | 245 | 264.74 | 0.1844 |
| 5 | PARCLOG, AGULHA, RETOENTB, MAOARTB | 246 | 266.96 | 0.1713 |
| 6 | Mod 5 + AGULHA*RETOENTB, AGULHA*MAOARTB, MAOARTB*RETOENTB, AGULHA*MAOARTB*RETOENTB | 242 | 263.60 | 0.1625 |
| 7 | Mod 5 + AGULHA*PARCLOG | 245 | 264.79 | 0.1838 |
| 8 | Mod 5 + RETOENTB*PARCLOG | 245 | 264.46 | 0.1876 |
| 9 | Mod 5 + MAOARTB*PARCLOG | 245 | 265.69 | 0.1737 |

feita pelos histogramas das Figuras 5.1 e 5.2. O novo modelo (2) foi ajustado com esta variável substituindo PARCEIRO. A melhora do ajuste é enorme, como mostra o $G^2$ do modelo. Outras transformações do número de parceiros foram tentadas, como a raiz quadrada, mas não lograram melhor resultado que o logaritmo.

O passo seguinte foi explorar a possibilidade de se descartar algumas das variáveis do modelo. Os testes $\chi^2$ para a hipótese do parâmetro ser zero são um bom indicativo das variáveis candidatas a serem retiradas. A decisão final sempre foi tomada com base em testes da razão de verossimilhança, pela diferença dos $G^2$ de dois modelos embutidos. A tabela dos $\chi^2$ para o modelo 2 (Tabela 5.3) indica 3 variáveis como passíveis de retirada: ARTRETB, MA-

```
Variable=PARCEIRO      Numero de parceiros

          Mean      48.54183  Med            15

                   Histogram                    #
975+*                                           4
   .
   .
   .
   .
   .*                                           1
   .
   .*                                           3
   .
   .*                                           5
   .*                                           1
   .*                                           2
   .*                                           2
   .*                                           9
   .*                                           3
   .**                                         20
   .*                                          11
   .****                                       46
   .*******                                    88
 25+*********************************************  582
    ----+----+----+----+----+----+----+----+----+
          * may represent up to 13 counts
```

Figura 5.1: Histograma da variável PARCEIRO.

```
Variable=PARCLOG Log do numero de parceiros + 1

          Mean       2.871413  Med       2.772589

                    Histogram                          #
6.75+**                                                5
    .***                                               9
    .******                                           16
    .***********                                      31
    .****************                                 48
    .**************                                   42
3.75+***************************                      76
    .************************************************ 134
    .*****************************                    87
    .********************************** .             98
    .*********************************                97
    .************************                          69
0.75+***********************                          65
     ----+----+----+----+----+----+----+----+----+
          * may represent up to 3 counts
```

Figura 5.2: Histograma da variável PARCLOG.

Tabela 5.3: Valores do teste $\chi^2$ para a hipótese de que os parâmetros do modelo 2 são zero.

| Variável | GL | QUI-QUADRADO | PROB |
|---|---|---|---|
| INTERCEPT | 1 | 96.80 | 0.0001 |
| PARCLOG | 1 | 58.17 | 0.0001 |
| AGULHA | 1 | 9.68 | 0.0019 |
| RETOENTB | 1 | 32.01 | 0.0001 |
| RETOPARB | 1 | 1.96 | 0.1618 |
| MAORETB | 1 | 3.63 | 0.0568 |
| ARTRETB | 1 | 1.61 | 0.2042 |

ORETB e RETOPARB. A variável MAORETB apresenta um p=0,057, no limite para um teste de significância de 5%. A variável ARTRETB já apresenta um p bem maior (p=0,204). Como estas 2 variáveis provavelmente descrevam um risco originado por traumatismo retal, tentamos substituí-las por apenas uma variável que foi gerada como sendo MAOARTB=ARTREB ou MAORETB. Isto é, esta nova variável assume valor 1 quando uma das duas é 1 e valor 0 quando as duas são 0. Um novo modelo foi ajustado incluindo esta variável (modelo 3). Em seguida retiramos as duas variáveis originais (modelo 4). Fazendo o teste, temos que $G^2_{(4)} - G^2_{(3)} = 4,20$ que tem distribuição $\chi^2$ com 2 graus de liberdade e portanto o teste é não significante a 5% – aceitamos que as duas variáveis retiradas sejam zero.

A próxima variável a ser examinada é RETOPARB, que pelo $\chi^2$, não está sendo importante no modelo. Assim o modelo 5 foi ajustado sem esta variável. O teste da razão de verossimilhança $G^2_{(5)} - G^2_{(4)} = 2,22$, que também é não significante a 5% e 1 grau de liberdade. Observando a Tabela 5.4 com os $\chi^2$ das variáveis vemos que não há mais nenhuma a ser retirada. Vencido este passo, estudaremos a conveniência da inclusão de algumas interações no modelo.

Para estudar as interações, foram ajustados os modelos de 6 a 9. As diferenças dos $G^2$ em relação ao modelo 5 mostram que em nenhum caso a diferença foi significante. A análise dos $\chi^2$ das novas variáveis também não indica que qualquer delas possa ter maior importância para o modelo. Assim,

Tabela 5.4: Valores do teste $\chi^2$ para a hipótese de que os parâmetros do modelo 5 são zero e dos coeficientes da regressão.

| Variável | $\hat{\beta}$ | $\exp(\hat{\beta})$ | GL | QUI-QUADRADO | PROB |
|---|---|---|---|---|---|
| INTERCEPT | -2.63 | 0.07 | 1 | 105.09 | 0.0001 |
| PARCLOG | 0.53 | 1.70 | 1 | 63.30 | 0.0001 |
| AGULHA | 1.22 | 3.39 | 1 | 10.21 | 0.0014 |
| RETOENTB | 1.21 | 3.35 | 1 | 36.74 | 0.0001 |
| MAOARTB | 0.52 | 1.68 | 1 | 6.55 | 0.0105 |

ficamos com o modelo 5 como o modelo mais simples que melhor se ajusta aos dados observados.

O modelo 5 parece concordar com o conhecimento atual sobre contaminação pelo vírus da AIDS. As variáveis que aparecem com maior peso no ajuste são aquelas que indicam intensidade de relações homossexuais, ou seja, o número de parceiros masculinos e o fato do entrevistado ser penetrado ou não pelos parceiros. Em seguida aparece a questão da partilha de agulhas para injeção endovenosa, fator que hoje ganha destaque em alguns centros. E finalmente um fator que deve representar um risco coadjuvante à relação homossexual pela ação traumática, que é o uso de artefatos ou da mão para introdução anal. Não se pode esquecer que estes dados foram obtidos em São Francisco, Califórnia, onde existe uma comunidade homossexual numerosa.

Em termos quantitativos, podemos avaliar o acréscimo de risco para cada um dos fatores em estudo através da exponencial dos parâmetros ajustados para cada fator. O número de parceiros é o principal fator de risco, apresentando um coeficiente de 0,53. Se lembrarmos que o número de parceiros em escala logaritmica vai até próximo de 7, a razão de odds para esta variável será de até $\exp(7 \times 0,53) = 40,85$ para um indivíduo com grande número de parceiros. Em seguida, AGULHA e RETOENTB apresentam uma razão de odds em torno de 3,4 para exposição. A menor razão de odds fica por conta de MAOARTB, com 1,68. Mais uma vez, a análise está de acordo com o conhecimento atual sobre AIDS, que revela que a promiscuidade representada pelo número de parceiros sexuais e pelo uso de agulhas compartilhadas por várias pessoas tem se constituído no fator mais importante

para a transmissão. Vemos também que o fato do indivíduo ser penetrado pelo parceiro constitui um risco adicional importante, enquanto que penetrar o parceiro não. Por fim, um risco coadjuvante é associado à prática da introdução anal de artefatos ou da mão ou punho, que deve aumentar o risco de contaminação por ação traumática, facilitando a entrada do vírus.

Outro aspecto interessante do ajuste diz respeito à transformação que se revelou apropriada para a variável PARCEIRO. A transformação logarítmica produz o efeito de expandir os valores baixos e comprimir fortemente os valores altos. A escala linear vai de 0 até próximo de 1000, sendo o meio da escala ocupado pelo valor 500. Os dados em escala logarítmica vão de 0 a próximo de 7, sendo 3,5 o meio da escala, ou 32 parceiros. Isto significa que, na verdade, o maior aumento do risco se dá com número intermediário de parceiros, sendo que o aumento do número de parceiros de 400 para 1000 (6 para 7 em log, aproximadamente) é menos importante que o aumento de 1 para 7 parceiros (0,7 para 2 em log, aproximadamente). Este fato é intensificado pelo fato de estarmos em escala logística.

Os programas utilizados para o ajuste dos modelos estão no Apêndice A e as listagens completas das saídas do computador estão no Apêndice B.

## 5.3 Diagnóstico

Após a escolha do modelo mais adequado aos dados, é importante que se faça uma avaliação crítica dos resultados para que se tenha certeza de que o modelo não está sendo perturbado por algumas das observações. Para isso, calculamos algumas quantidades descritas no capítulo anterior, na seção 4.3: os $g_i$, os $\chi_i$, os $h_{ii}$ e os $c_i^1$. Os programas utilizados para o cálculo desses elementos estão no Apêndice A. Os resultados são apresentados em forma de gráficos. A Figura 5.3 apresenta o gráfico dos $g_i$ contra o número da observação e a Figura 5.4, uma análise descritiva dos $g_i$. A Figura 5.5 apresenta o gráfico dos elementos da diagonal de $\mathbf{H}$ contra o número da observação. A Figura 5.6 apresenta o gráfico dos $c_i^1$ contra o número da observação.

O gráfico que apresenta os componentes da estatística da equação de verossimilhança ($g_i$) mostra que não há observações muito distanciadas das outras. Quatro observações se situam um pouco mais distantes (93, 103,

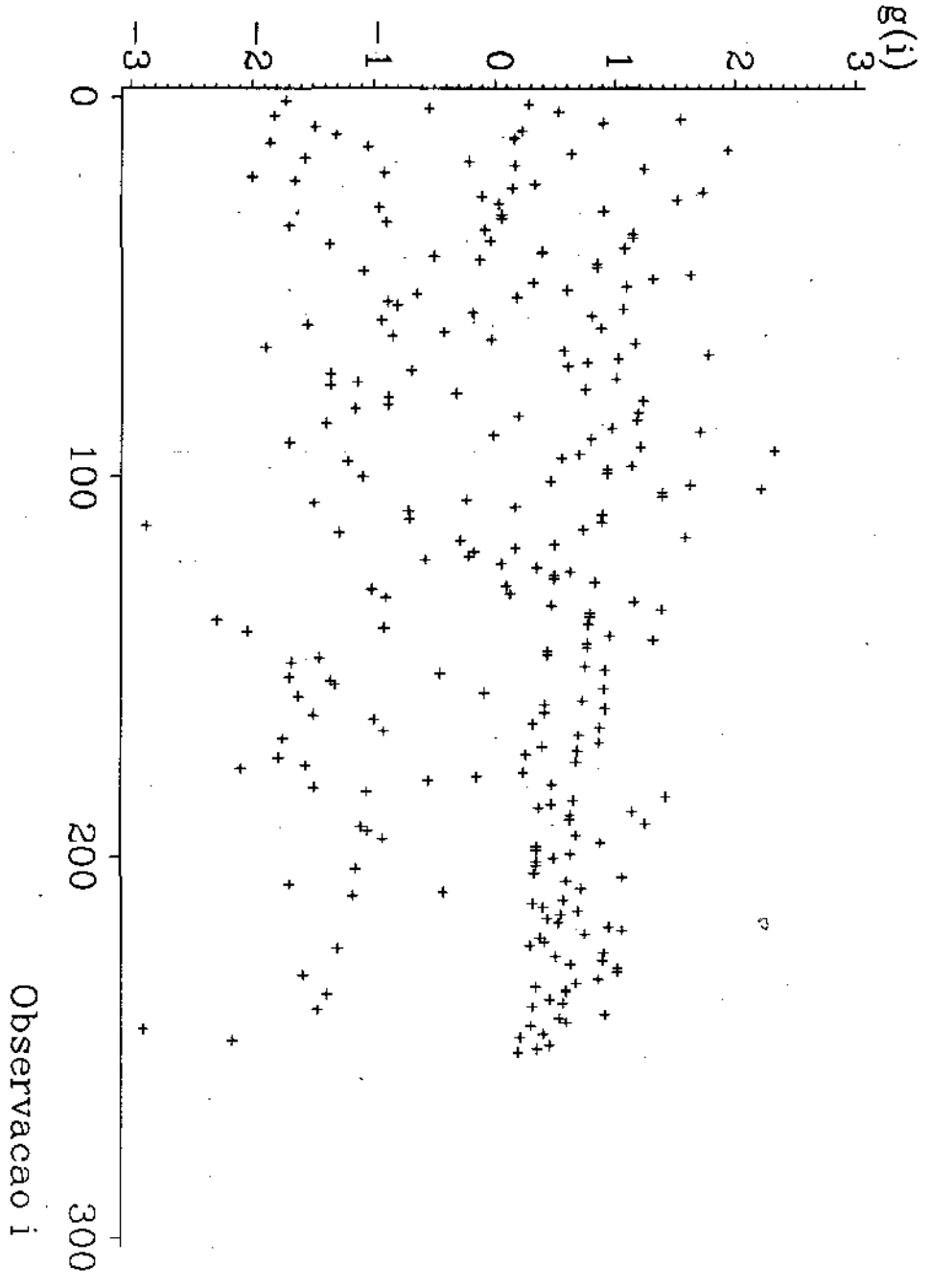


Figura 5.3: Gráfico de $g_i \times i$.

```
                    UNIVARIATE PROCEDURE

Variable=GI

          N              251  Sum Wgts       251
          Mean       0.066333  Med        0.353579
          Std Dev    1.031218  Variance   1.063411
          Skewness  -0.57852   Kurtosis  -0.44905
          USS         266.957   CSS        265.8526
          CV         1554.618   Std Mean    0.06509
          T:Mean=0   1.019092   Prob>|T|    0.3091

                        Extremes
            Lowest     Obs      Highest    Obs
           -2.89117(     245) 1.738377(     25)
           -2.87306(     113) 1.779321(     68)
           -2.28492(     138) 1.941333(     14)
           -2.15783(     248) 2.222968(    103)
           -2.08573(     177) 2.327207(     93)

                  Histogram                  #
 2.25+**                                     2
     .*****                                  9
     .**************                        27
     .**********************************    65
     .*****************************         56
-0.25+*********                              18
     .***********                           22
     .**************                        28
     .*********                             18
     .**                                     4
-2.75+**                                     2
      ----+----+----+----+----+----+---
      * may represent up to 2 counts
```

Figura 5.4: Análise descritiva dos componentes de $G^2$.

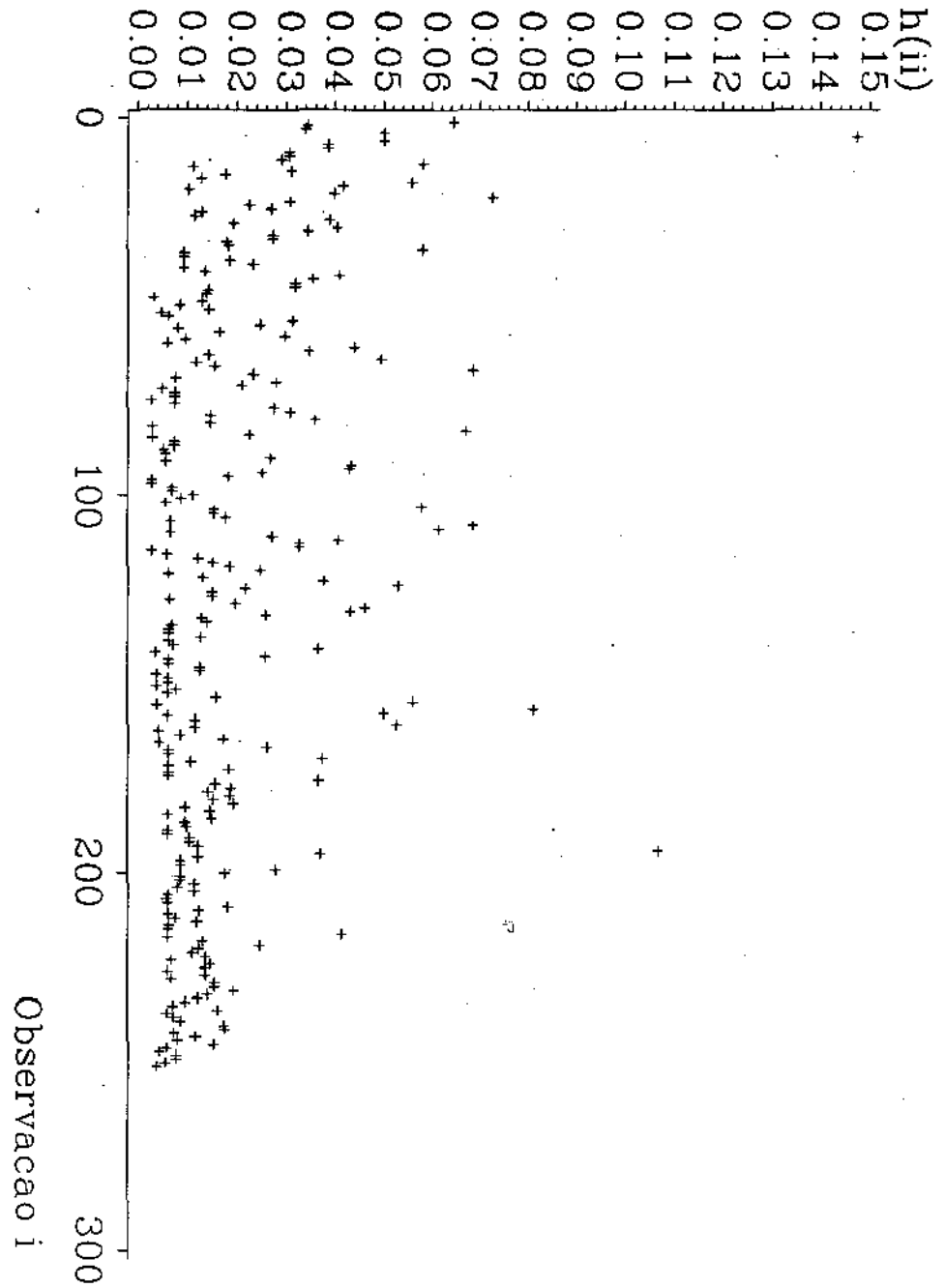


Figura 5.5: Gráfico dos $h_{ii} \times i$.

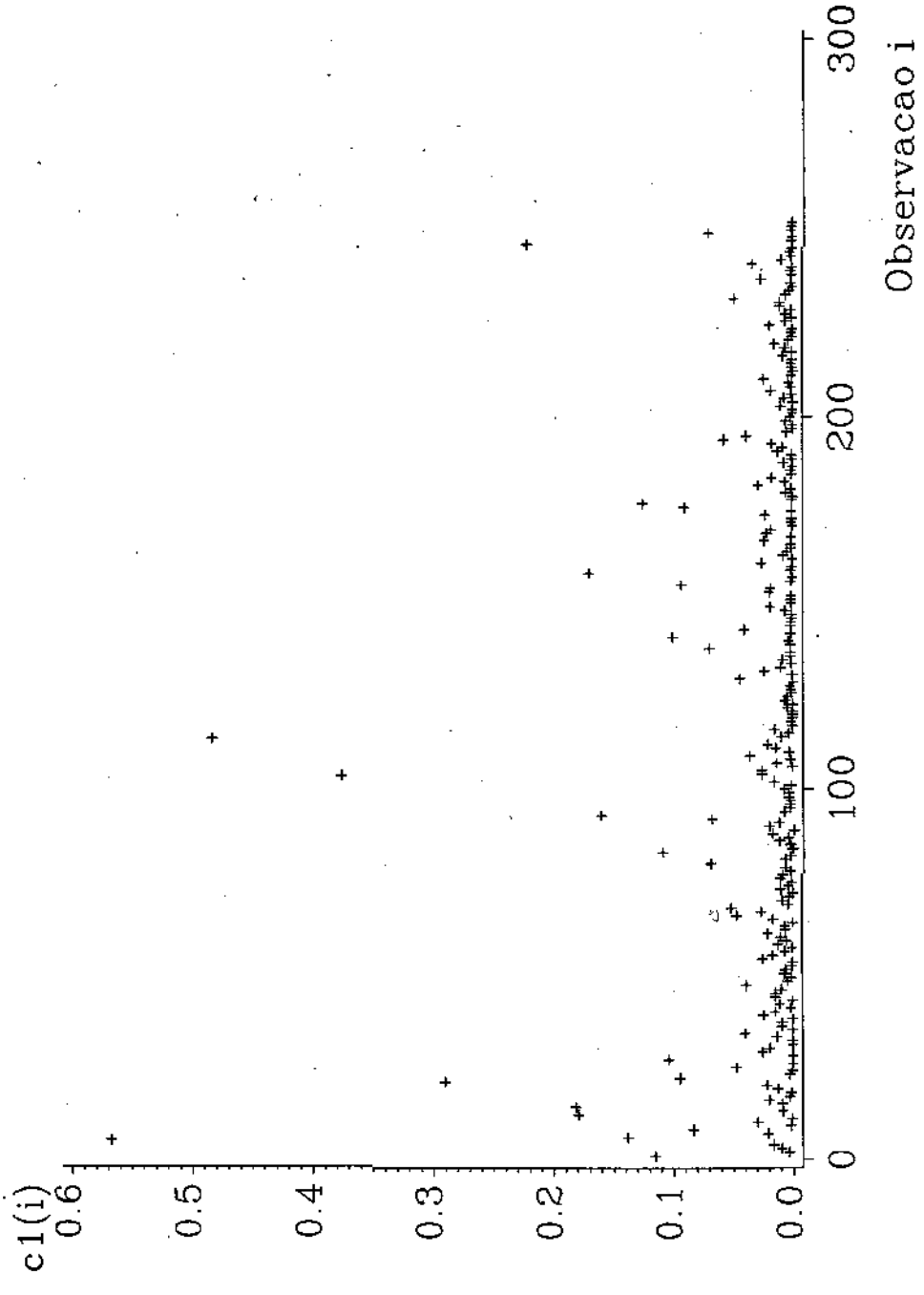


Figura 5.6: Gráfico de $c_i^1 \times i$.

113 e 245), mas não parecem constituir problema. Uma discreta tendência descendente aparece neste gráfico, indicando que, talvez, uma outra transformação pudesse melhorar um pouco mais o ajuste. Aparentemente há uma discreta tendência de resíduos positivos para valores baixos de PARCLOG e negativos para valores altos. Uma análise descritiva dos $d_i$ aparece na Figura 5.4 e mostra que eles têm uma distuibuição próxima da normal com média 0 e variância 1. Veja o gráfico probabilístico normal no Apêndice B.

O gráfico dos elementos da diagonal de $\mathbf{H}$ ($h_{ii}$) apresenta 2 observações destacadas, a observação 5 e a 194. O gráfico dos $c_i^1$ apresenta 3 observações destacadas: 5, 103 e 113. A observação 5 se repete nos 2 gráficos, apresentando-se como a mais influente no ajuste. Vejamos o valor das variáveis em estudo para estas observações:

| OBS | PARCLOG | AGUALHA | RETOENTB | MAOARTB | TOALL | CASOS | _PRED_ |
|---|---|---|---|---|---|---|---|
| 5 | 0.69315 | 0 | 1 | 0 | 26 | 3 | 0.25916 |
| 103 | 3.04452 | 0 | 0 | 0 | 10 | 6 | 0.26493 |
| 113 | 3.04452 | 1 | 1 | 1 | 2 | 0 | 0.87301 |
| 194 | 4.61512 | 0 | 1 | 0 | 20 | 16 | 0.73446 |

A observação 5 apresenta um número baixo de parceiros, e valor 1 apenas para RETOENTB. A observação 103 apresenta um número médio de parceiros e 0 para as outras variáveis e a observação 113, também um número médio de parceiros e 1 para as outras variáveis. A observação 194 apresenta um número relativamente alto de parceiros e 1 apenas para RETOENTB. Não é difícil perceber que são todas pontos extremos no espaço de desenho, o que confere esta característica de grande influência. O cuidado que deve se ter com estas observações é de se certificar da correção das respostas aos

questionários, para evitar que pequenos erros possam provocar grandes alterações no modelo. Por outro lado, se os dados estão corretos, não é lícito retirar estas observações do conjunto de dados, visto que, por sua influência no ajuste são importantes e carregam muita informação proporcionalmente às outras observações visto que estão relativamente solitárias no espaço de desenho, trazem informação sobre grupos de indivíduos que as outras variáveis não trazem. Como afirma Pregibon (1981 [16]), o intuito desta análise não é criar modelos utópicos e artificiais pela retirada dos pontos que se mostram mal ajustados ou influentes, mas sim identificar os pontos que apresentam um impacto substancial no ajuste.

# Referências Bibliográficas


[1] BICKEL, P. J. & DOKSUM, K. A. (1977). *Mathematical Statistics.* San Francisco, Holden-Day Inc.

[2] BISHOP, Y. M. M.; FIENBERG, S. E. & HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, The MIT Press.

[3] BRESLOW, N. E. & DAY, N. E. (1980). *Statistical Methods in Cancer Research. Vol. 1.* Lyon, International Agency for Research on Cancer.

[4] CORNELL, R. G. (1982). Biostatistics and Epidemiology. *Communications in Statistics A. Theory and Methods. 11*: 445–448.

[5] COX, D. R. (1970). *The analysis of binary data.* London, Methuen & Co.

[6] FIENBERG, S. (1977). *The Analysis of Cross-classified Categorical Data.* Cambridge (MA), The MIT Press.

[7] GRIZZLE, J. E.; STARMER, C. F. & KOCH, G. G. (1969). Analysis of categorical data by linear models. *Biometrics 25*: 489-504.

[8] KAY, R. & LITTLE, S. (1986). Assessing the fit of the logistic model: a case study of children with haemolitic anaemic syndrome. *Applied Statistics 35*: 16–30.

[9] KLEINBAUM, D. G.; KUPPER, L. L. & CHAMBLESS, L. E. (1982). Logistic regression analysis of epidemiologic data: theory and practice. *Communications in Statistics A 11*: 485–547.

[10] KLEINBAUM, D. G.; KUPPER, L. L. & MORGENSTERN, H. (1982). *Epidemiologic Research*. Belmont, Lifetime Learning Publications.

[11] LANDWEHR, J. M.; PREGIBON, D. & SHOEMAKER, A. C. (1984). Graphical methods for assessing logistic regression models. *JASA 79*: 61–83.

[12] MIETTINEN, O. (1976). Estimability and estimation in case-referent studies. *American Journal of Epidemiology 103*: 226–235.

[13] MIETTINEN, O. & COOK, F. (1981). Confounding: essence and detection. *American Journal of Epidemiology 114*: 593–603.

[14] MORGENSTERN, H.; KLEINBAUM, D. G. & KUPPER, L. L. (1980). Measures of disease incidence used in epidemiologic research. *International Journal of Epidemiology 9*: 97–104.

[15] MORRISON, A. S. (1979). Sequential pathogenic components of rates. *American Journal of Epidemiology 109*: 709–718.

[16] PREGIBON, D. (1981). Logistic regression diagnostics. *The Annals of Statistics 9*:705–724.

[17] PRENTICE, R. L. & PYKE, R. (1979). Logistic disease incidence models in case-control studies. *Biometrika 66*: 403–411.

[18] ROSE, G. & BARKER, D. J. P. (1979). *Epidemiology for the uninitiated*. London, British Medical Association.

[19] ROTHMAN, K. J. (1986). *Modern Epidemiology*. Boston, Little, Brown & Co.

[20] WILLIAMS, D. A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics 36*: 181–191.

# Apêndice A
# Listagem dos programas

Neste apêndice apresentamos as listagens comentadas dos programas utilizados para a análise do conjunto de dados apresentada no Capítulo 5. Toda a parte computacional foi realizada com o SAS, quase todo o tempo rodando em um microcomputador do tipo PC–XT. Um IBM 3090 foi utilizado eventualmente com o intuito apenas de reduzir o tempo de processamento, visto que toda a análise pode ser realizada no PC.

Para ajustar um modelo de regressão logística através do SAS, utiliza-se o procedimento CATMOD. Este procedimento é genérico, permitindo o ajuste de vários modelos para dados categóricos, como o modelo log-linear e o modelo logístico ajustado por quadrados mínimos ponderados. Para o ajuste por Máxima Verossimilhança, as opções ML e NOGLS devem ser utilizadas. O programa a seguir ajusta o modelo 1 da Tabela 5.2:

```
title 'Modelo 1';
proc catmod data=baseline;
   population parceiro agulha retoentb retoparb maoretb artretb;
   direct     parceiro agulha retoentb retoparb maoretb artretb;
   model hiv=parceiro agulha retoentb retoparb maoretb artretb
             / nogls ml noprofile;
run; quit;
```

A declaração POPULATION define quais variáveis servirão para a divisão dos dados em populações distintas. Se esta declaração for omitida, as populações serão criadas a partir das variáveis citadas na declaração MODEL. Isto pode não ser conveniente quando modelos seqüenciais estão sendo ajustados e queremos sempre comparar com o mesmo modelo saturado. A declaração DIRECT faz com que as variáveis incluídas no modelo sejam coloca-

das na matriz de desenho diretamente, sem a criação de variáveis indicadoras. Assim, temos total controle da matriz de desenho. Se nosso exemplo tivesse variáveis com mais de 2 níveis, teríamos que criar as variáveis indicadoras necessárias.

A seguir, criamos um novo conjunto de dados que contém as variáveis do anterior mais a variável PARCLOG:

```
data baselin2;
   set baselin2;
   parclog=log(parceiro+1);
   if parceiro=. or agulha=. or retoentb=. or retoparb=. or
       maoretb=. or artretb=. then delete;
run;
```

e aproveitamos para retirar do arquivo as observações que contêm alguma variável com dados registrados como faltante (*missing*), visto que estas observações não são utilizadas no ajuste do modelo.

A seguir ajustamos o modelo 2 e criamos a variável MAOARTB:

```
proc catmod data=baselin2;
   population parceiro agulha retoentb retoparb maoretb artretb;
   direct     parceiro agulha retoentb retoparb maoretb artretb
              parclog;
   model hiv=parclog agulha retoentb retoparb maoretb artretb
         / nogls ml noprofile;
run; quit;

data baselin2;
   set baselin2;
   maoartb=(maoretb=1 or artretb=1);
   output;
run;
```

E em seguida ajustamos os modelos de 3 a 9:

```
title 'Modelo 3';
proc catmod data=baselin2;
   population parceiro agulha retoentb retoparb maoretb artretb;
   direct     parceiro agulha retoentb retoparb maoretb artretb
```

```
              parclog maoartb;
   model hiv=parclog agulha retoentb retoparb maoretb artretb
             maoartb / nogls ml noprofile;
run; quit;

title 'Modelo 4';
proc catmod data=baselin2;
   population parceiro agulha retoentb retoparb maoretb artretb;
   direct     parceiro agulha retoentb retoparb maoretb artretb
               parclog maoartb;
   model hiv=parclog agulha retoentb retoparb maoartb
          / nogls ml noprofile;
run; quit;

title 'Modelo 5';
proc catmod data=baselin2;
   population parceiro agulha retoentb retoparb maoretb artretb;
   direct     parceiro agulha retoentb retoparb maoretb artretb
               parclog maoartb;
   model hiv=parclog agulha retoentb maoartb
          / nogls ml noprofile;
run; quit;

title 'Modelo 6';
proc catmod data=baselin2;
   population parceiro agulha retoentb retoparb maoretb artretb;
   direct     parceiro agulha retoentb retoparb maoretb artretb
               parclog maoartb;
   model hiv=parclog agulha|retoentb|maoartb
          / nogls ml noprofile;
run; quit;

title 'Modelo 7';
proc catmod data=baselin2;
   population parceiro agulha retoentb retoparb maoretb artretb;
   direct     parceiro agulha retoentb retoparb maoretb artretb
               parclog maoartb;
```

```
   model hiv=parclog agulha retoentb maoartb agulha*parclog
         / nogls ml noprofile;
run; quit;

title 'Modelo 8';
proc catmod data=baselin2;
   population parceiro agulha retoentb retoparb maoretb artretb;
   direct     parceiro agulha retoentb retoparb maoretb artretb
               parclog maoartb;
   model hiv=parclog agulha retoentb maoartb retoentb*parclog
         / nogls ml noprofile;
run; quit;

title 'Modelo 9';
proc catmod data=baselin2;
   population parceiro agulha retoentb retoparb maoretb artretb;
   direct     parceiro agulha retoentb retoparb maoretb artretb
               parclog maoartb;
   model hiv=parclog agulha retoentb maoartb maoartb*parclog
         / nogls ml noprofile;
run; quit;
```

Agora apresentamos os programas utilizados para calcular as medidas de diagnóstico. Como o conjunto de dados original apresentava uma linha para cada das 777 observações válidas e temos apenas 251 populações (caselas), vamos compactar o arquivo de modo a termos apenas uma linha para cada população. Isto possibilita que as probabilidades ajustadas sejam associadas a cada população sem dificuldade.

```
proc summary nway data=baselin2;
   class parclog agulha retoentb retoparb maoretb artretb
         maoartb;
   var hiv;
   output out=basesumm n=total sum=casos;
run;
```

Para que os dados nesta forma possam ser utilizados pelo CATMOD, é necessário que a variável resposta binária seja recriada e uma variável com

pesos (número de casos ou controles) seja calculada a partir do número de casos e do total de cada população:

```
data basesum2;
   set basesumm;
   hiv=0; peso=total-casos; output;
   hiv=1; peso=casos; output;
run;
```

Agora estamos prontos para rodar o modelo escolhido como melhor e pedir que as probabilidades ajustadas sejam gravadas em um arquivo (BASEPRED). As variáveis da declaração POPULATION devem ser exatamente as mesmas e na mesma ordem das colocadas na declaração CLASS do procedimento SUMMARY, de forma que as populações formadas estejam na mesma ordem.

```
proc catmod data=basesum2;
   weight peso;
   population parclog agulha retoentb retoparb maoretb artretb
              maoartb;
   direct     parclog agulha retoentb retoparb maoretb artretb
              maoartb;
   response / out=basepred;
   model hiv=parclog agulha retoentb maoartb
         / nogls ml noprofile noiter;
run; quit;
```

O arquivo criado tem mais informações do que desejamos, assim vamos apagar as linhas que contêm informações que não dizem respeito à probabilidade de adoecer e conservar apenas as variáveis que nos interessam (_PRED_ e _SAMPLE_):

```
data basepred;
   set basepred;
   if _number_^=2 then delete;
   keep _sample_ _pred_;
run;
```

Agora vamos associar as probabilidades ajustadas pelo modelo 5 ao conjunto de dados compactado, de forma que tenhamos todas as informações necessárias para calcular as medidas de interesse:

```
data basesumm;
   merge basesumm basepred;
run;
```

Finalmente, calculamos as medidas $g_i$ e $\chi_i$ e algumas outras variáveis auxiliares:

```
data basesum3;
   set basesumm;
   freqobs=casos/total;
   logverh0=casos*log(_pred_) + (total-casos)*log(1-_pred_);
   if casos=0 or casos=total then logverh1=0;
   else logverh1=(casos*log(freqobs) +
                                  (total-casos)*log(1-freqobs));
   g2i=-2*(logverh0-logverh1);
   gi=(-1)**(freqobs<_pred_)*sqrt(g2i);
   vii=total*_pred_*(1-_pred_);
   xi=(casos-(total*_pred_))/sqrt(vii);
   output;
run;
```

Para calcular os elementos da diagonal da matriz de projeção, é muito mais fácil utilizar um pacote que faça cálculos matriciais. Mesmo assim, a matriz de covariância, na maioria dos casos, não caberá na memória do computador. Assim, usamos um artifício para calcular apenas a diagonal da matriz de projeção e não a matriz inteira. Isto é muito mais eficiente em termos de utilização de memória e de tempo de processamento. No nosso caso utilizamos o IML, linguagem acoplada ao SAS.

```
proc iml wrksize=90;
use basesum3 var{_pred_ total casos parclog retoentb agulha
                 maoartb vii};
read all var{parclog agulha retoentb maoartb} into x;
read all var{total} into freq;
read all var{vii} into v;
read all var{_pred_} into pc;
close basesum3;

x=j(nrow(x),1)||x;
```

```
v=sqrt(v);
xlv=j(ncol(x),nrow(x));
xl=x';
do i=1 to nrow(xlv);
   do j=1 to ncol(xlv);
      xlv[i,j]=xl[i,j]*v[j,1];
   end;
end;
s=inv(xlv*xlv');
hii=j(nrow(x),1);
do i=1 to nrow(x);
   hii[i,1]=xlv[,i]'*s*xlv[,i];
end;

create diagh from hii (|colname=\{hii\}|);
edit diagh; setout diagh;
append from hii;
close diagh;
quit;
```

Os elementos da diagonal de H foram colocados num arquivo de nome DIAGH, que agora pode ser mesclado ao nosso arquivo principal:

```
data basesum3;
   merge basesum3 diagh;
run;
```

Finalizando, calculamos os valores dos $c_i^1$:

```
data basesum3;
   set basesum3;
   c1i=(xi**2*hii)/((1-hii)**2);
run;
```

# Apêndice B
# Listagem das saídas de computador

Modelo 1

CATMOD PROCEDURE

| | |
|---|---|
| RESPONSE: HIV | RESPONSE LEVELS (R)= 2 |
| WEIGHT VARIABLE: | POPULATIONS (S)= 251 |
| DATA SET: BASELINE | TOTAL FREQUENCY (N)= 777 |
| | OBSERVATIONS (OBS)= 777 |

ANALYSIS OF VARIANCE TABLE

| SOURCE | DF | CHI-SQUARE | PROB |
|---|---|---|---|
| INTERCEPT | 1 | 59.16 | 0.0001 |
| PARCEIRO | 1 | 15.39 | 0.0001 |
| AGULHA | 1 | 11.34 | 0.0008 |
| RETOENTB | 1 | 32.33 | 0.0001 |
| RETOPARB | 1 | 5.87 | 0.0154 |
| MAORETB | 1 | 4.62 | 0.0316 |
| ARTRETB | 1 | 3.16 | 0.0754 |
| LIKELIHOOD RATIO | 244 | 306.31 | 0.0041 |

ANALYSIS OF INDIVIDUAL PARAMETERS

| EFFECT | PARAMETER | ESTIMATE | STANDARD ERROR | CHI-SQUARE | PROB |
|---|---|---|---|---|---|
| INTERCEPT | 1 | 1.7014 | 0.22121 | 59.16 | 0.0001 |
| PARCEIRO | 2 | -.005051 | .0012876 | 15.39 | 0.0001 |
| AGULHA | 3 | -1.26181 | 0.374659 | 11.34 | 0.0008 |
| RETOENTB | 4 | -1.1333 | 0.199327 | 32.33 | 0.0001 |
| RETOPARB | 5 | -0.512 | 0.21125 | 5.87 | 0.0154 |
| MAORETB | 6 | -.836819 | 0.389409 | 4.62 | 0.0316 |
| ARTRETB | 7 | -0.36805 | 0.20701 | 3.16 | 0.0754 |

Modelo 2

CATMOD PROCEDURE

| | | |
|---|---|---|
| RESPONSE: HIV | RESPONSE LEVELS (R)= | 2 |
| WEIGHT VARIABLE: | POPULATIONS (S)= | 251 |
| DATA SET: BASELINE | TOTAL FREQUENCY (N)= | 777 |
| | OBSERVATIONS (OBS)= | 777 |

ANALYSIS OF VARIANCE TABLE

| SOURCE | DF | CHI-SQUARE | PROB |
|---|---|---|---|
| INTERCEPT | 1 | 96.80 | 0.0001 |
| PARCLOG | 1 | 58.17 | 0.0001 |
| AGULHA | 1 | 9.68 | 0.0019 |
| RETOENTB | 1 | 32.01 | 0.0001 |
| RETOPARB | 1 | 1.96 | 0.1618 |
| MAORETB | 1 | 3.63 | 0.0568 |
| ARTRETB | 1 | 1.61 | 0.2042 |
| LIKELIHOOD RATIO | 244 | 263.57 | 0.1859 |

ANALYSIS OF INDIVIDUAL PARAMETERS

| EFFECT | PARAMETER | ESTIMATE | STANDARD ERROR | CHI-SQUARE | PROB |
|---|---|---|---|---|---|
| INTERCEPT | 1 | 2.77568 | 0.282123 | 96.80 | 0.0001 |
| PARCLOG | 2 | -.512095 | .0671446 | 58.17 | 0.0001 |
| AGULHA | 3 | -1.19638 | 0.384524 | 9.68 | 0.0019 |
| RETOENTB | 4 | -1.15974 | 0.204973 | 32.01 | 0.0001 |
| RETOPARB | 5 | -.307062 | 0.21947 | 1.96 | 0.1618 |
| MAORETB | 6 | -.757123 | 0.397469 | 3.63 | 0.0568 |
| ARTRETB | 7 | -.272361 | 0.214494 | 1.61 | 0.2042 |

Modelo 3

CATMOD PROCEDURE

| | |
|---|---|
| RESPONSE: HIV | RESPONSE LEVELS (R)= 2 |
| WEIGHT VARIABLE: | POPULATIONS (S)= 251 |
| DATA SET: BASELINE | TOTAL FREQUENCY (N)= 777 |
| | OBSERVATIONS (OBS)= 777 |

ANALYSIS OF VARIANCE TABLE

| SOURCE | DF | CHI-SQUARE | PROB |
|---|---|---|---|
| INTERCEPT | 1 | 96.70 | 0.0001 |
| PARCLOG | 1 | 57.53 | 0.0001 |
| AGULHA | 1 | 9.88 | 0.0017 |
| RETOENTB | 1 | 31.43 | 0.0001 |
| RETOPARB | 1 | 1.90 | 0.1680 |
| MAORETB | 1 | 0.64 | 0.4252 |

| | | | |
|---|---|---|---|
| ARTRETB | 1 | 1.52 | 0.2182 |
| MAOARTB | 1 | 2.32 | 0.1279 |
| LIKELIHOOD RATIO | 243 | 260.54 | 0.2098 |

ANALYSIS OF INDIVIDUAL PARAMETERS

| EFFECT | PARAMETER | ESTIMATE | STANDARD ERROR | CHI-SQUARE | PROB |
|---|---|---|---|---|---|
| INTERCEPT | 1 | 2.7756 | 0.282262 | 96.70 | 0.0001 |
| PARCLOG | 2 | -.509212 | .0671325 | 57.53 | 0.0001 |
| AGULHA | 3 | -1.20869 | 0.38444 | 9.88 | 0.0017 |
| RETOENTB | 4 | -1.15238 | 0.205564 | 31.43 | 0.0001 |
| RETOPARB | 5 | -.302995 | 0.219759 | 1.90 | 0.1680 |
| MAORETB | 6 | -.356388 | 0.446966 | 0.64 | 0.4252 |
| ARTRETB | 7 | 1.40796 | 1.14339 | 1.52 | 0.2182 |
| MAOARTB | 8 | -1.77043 | 1.16277 | 2.32 | 0.1279 |

**Modelo 4**

CATMOD PROCEDURE

| | |
|---|---|
| RESPONSE: HIV | RESPONSE LEVELS (R)= 2 |
| WEIGHT VARIABLE: | POPULATIONS (S)= 251 |
| DATA SET: BASELINE | TOTAL FREQUENCY (N)= 777 |
| | OBSERVATIONS (OBS)= 777 |

ANALYSIS OF VARIANCE TABLE

| SOURCE | DF | CHI-SQUARE | PROB |
|---|---|---|---|
| INTERCEPT | 1 | 97.62 | 0.0001 |
| PARCLOG | 1 | 57.89 | 0.0001 |

| | | | |
|---|---|---|---|
| AGULHA | 1 | 9.96 | 0.0016 |
| RETOENTB | 1 | 30.94 | 0.0001 |
| RETOPARB | 1 | 2.20 | 0.1377 |
| MAOARTB | 1 | 6.45 | 0.0111 |
| LIKELIHOOD RATIO | 245 | 264.74 | 0.1844 |

ANALYSIS OF INDIVIDUAL PARAMETERS

| EFFECT | PARAMETER | ESTIMATE | STANDARD ERROR | CHI-SQUARE | PROB |
|---|---|---|---|---|---|
| INTERCEPT | 1 | 2.78753 | 0.282132 | 97.62 | 0.0001 |
| PARCLOG | 2 | -.510359 | .0670796 | 57.89 | 0.0001 |
| AGULHA | 3 | -1.20716 | 0.382533 | 9.96 | 0.0016 |
| RETOENTB | 4 | -1.13971 | 0.204912 | 30.94 | 0.0001 |
| RETOPARB | 5 | -.325416 | 0.219224 | 2.20 | 0.1377 |
| MAOARTB | 6 | -.512237 | 0.201677 | 6.45 | 0.0111 |

**Modelo 5**

CATMOD PROCEDURE

| | |
|---|---|
| RESPONSE: HIV | RESPONSE LEVELS (R)= 2 |
| WEIGHT VARIABLE: | POPULATIONS (S)= 251 |
| DATA SET: BASELINE | TOTAL FREQUENCY (N)= 777 |
| | OBSERVATIONS (OBS)= 777 |

ANALYSIS OF VARIANCE TABLE

| SOURCE | DF | CHI-SQUARE | PROB |
|---|---|---|---|
| INTERCEPT | 1 | 105.09 | 0.0001 |
| PARCLOG | 1 | 63.30 | 0.0001 |

| | | | |
|---|---|---|---|
| AGULHA | 1 | 10.21 | 0.0014 |
| RETOENTB | 1 | 36.74 | 0.0001 |
| MAOARTB | 1 | 6.55 | 0.0105 |
| LIKELIHOOD RATIO | 246 | 266.96 | 0.1713 |

ANALYSIS OF INDIVIDUAL PARAMETERS

| EFFECT | PARAMETER | ESTIMATE | STANDARD ERROR | CHI-SQUARE | PROB |
|---|---|---|---|---|---|
| INTERCEPT | 1 | 2.62562 | 0.256122 | 105.09 | 0.0001 |
| PARCLOG | 2 | -.527215 | .0662666 | 63.30 | 0.0001 |
| AGULHA | 3 | -1.223 | 0.382698 | 10.21 | 0.0014 |
| RETOENTB | 4 | -1.20983 | 0.199586 | 36.74 | 0.0001 |
| MAOARTB | 5 | -.515477 | 0.201451 | 6.55 | 0.0105 |

Modelo 6

CATMOD PROCEDURE

| | |
|---|---|
| RESPONSE: HIV | RESPONSE LEVELS (R)= 2 |
| WEIGHT VARIABLE: | POPULATIONS (S)= 251 |
| DATA SET: BASELINE | TOTAL FREQUENCY (N)= 777 |
| | OBSERVATIONS (OBS)= 777 |

ANALYSIS OF VARIANCE TABLE

| SOURCE | DF | CHI-SQUARE | PROB |
|---|---|---|---|
| INTERCEPT | 1 | 101.08 | 0.0001 |
| PARCLOG | 1 | 63.18 | 0.0001 |
| AGULHA | 1 | 0.01 | 0.9423 |
| RETOENTB | 1 | 33.81 | 0.0001 |

| | | | |
|---|---|---|---|
| AGULHA*RETOENTB | 1 | 1.51 | 0.2196 |
| MAOARTB | 1 | 2.30 | 0.1291 |
| AGULHA*MAOARTB | 1 | 0.00 | 0.9765 |
| RETOENTB*MAOARTB | 1 | 0.74 | 0.3907 |
| AGULHA*RETOENTB*MAOARTB | 1 | 0.00 | 0.9757 |
| LIKELIHOOD RATIO | 242 | 263.60 | 0.1625 |

ANALYSIS OF INDIVIDUAL PARAMETERS

| EFFECT | PARAMETER | ESTIMATE | STANDARD ERROR | CHI-SQUARE | PROB |
|---|---|---|---|---|---|
| INTERCEPT | 1 | 2.64293 | 0.262873 | 101.08 | 0.0001 |
| PARCLOG | 2 | -0.52927 | .0665849 | 63.18 | 0.0001 |
| AGULHA | 3 | .0881907 | 1.21909 | 0.01 | 0.9423 |
| RETOENTB | 4 | -1.22596 | 0.21085 | 33.81 | 0.0001 |
| AGULHA*RETOENTB | 5 | -1.69459 | 1.38047 | 1.51 | 0.2196 |
| MAOARTB | 6 | -1.14695 | 0.755767 | 2.30 | 0.1291 |
| AGULHA*MAOARTB | 7 | -13.1757 | 446.506 | 0.00 | 0.9765 |
| RETOENTB*MAOARTB | 8 | 0.674881 | 0.786229 | 0.74 | 0.3907 |
| AGULHA*RETOENTB*MAOARTB | 9 | 13.6203 | 446.507 | 0.00 | 0.9757 |

Modelo 7

CATMOD PROCEDURE

| | |
|---|---|
| RESPONSE: HIV | RESPONSE LEVELS (R)= 2 |
| WEIGHT VARIABLE: | POPULATIONS (S)= 251 |
| DATA SET: BASELINE | TOTAL FREQUENCY (N)= 777 |
| | OBSERVATIONS (OBS)= 777 |

ANALYSIS OF VARIANCE TABLE

| SOURCE | DF | CHI-SQUARE | PROB |
|---|---|---|---|
| INTERCEPT | 1 | 98.23 | 0.0001 |
| PARCLOG | 1 | 55.78 | 0.0001 |
| AGULHA | 1 | 0.01 | 0.9323 |
| RETOENTB | 1 | 36.67 | 0.0001 |
| MAOARTB | 1 | 6.61 | 0.0101 |
| AGULHA*PARCLOG | 1 | 1.87 | 0.1715 |
| LIKELIHOOD RATIO | 245 | 264.79 | 0.1838 |

ANALYSIS OF INDIVIDUAL PARAMETERS

| EFFECT | PARAMETER | ESTIMATE | STANDARD ERROR | CHI-SQUARE | PROB |
|---|---|---|---|---|---|
| INTERCEPT | 1 | 2.56199 | 0.258499 | 98.23 | 0.0001 |
| PARCLOG | 2 | -.504704 | .0675759 | 55.78 | 0.0001 |
| AGULHA | 3 | .0841235 | 0.990356 | 0.01 | 0.9323 |
| RETOENTB | 4 | -1.20826 | 0.199521 | 36.67 | 0.0001 |
| MAOARTB | 5 | -.518885 | 0.201751 | 6.61 | 0.0101 |
| AGULHA*PARCLOG | 6 | -0.47175 | 0.34496 | 1.87 | 0.1715 |

Modelo 8

CATMOD PROCEDURE

| | |
|---|---|
| RESPONSE: HIV | RESPONSE LEVELS (R)= 2 |
| WEIGHT VARIABLE: | POPULATIONS (S)= 251 |
| DATA SET: BASELINE | TOTAL FREQUENCY (N)= 777 |
| | OBSERVATIONS (OBS)= 777 |

ANALYSIS OF VARIANCE TABLE

| SOURCE | DF | CHI-SQUARE | PROB |
|---|---|---|---|
| INTERCEPT | 1 | 29.14 | 0.0001 |
| PARCLOG | 1 | 7.49 | 0.0062 |
| AGULHA | 1 | 10.07 | 0.0015 |
| RETOENTB | 1 | 1.38 | 0.2400 |
| MAOARTB | 1 | 6.08 | 0.0137 |
| RETOENTB*PARCLOG | 1 | 2.53 | 0.1114 |
| LIKELIHOOD RATIO | 245 | 264.46 | 0.1876 |

ANALYSIS OF INDIVIDUAL PARAMETERS

| EFFECT | PARAMETER | ESTIMATE | STANDARD ERROR | CHI-SQUARE | PROB |
|---|---|---|---|---|---|
| INTERCEPT | 1 | 2.12312 | 0.393331 | 29.14 | 0.0001 |
| PARCLOG | 2 | -.350145 | 0.127907 | 7.49 | 0.0062 |
| AGULHA | 3 | -1.2179 | 0.38379 | 10.07 | 0.0015 |
| RETOENTB | 4 | -.539562 | 0.459197 | 1.38 | 0.2400 |
| MAOARTB | 5 | -.501227 | 0.203269 | 6.08 | 0.0137 |
| RETOENTB*PARCLOG | 6 | -.238802 | 0.149993 | 2.53 | 0.1114 |

Modelo 9

CATMOD PROCEDURE

| | |
|---|---|
| RESPONSE: HIV | RESPONSE LEVELS (R)= 2 |
| WEIGHT VARIABLE: | POPULATIONS (S)= 251 |
| DATA SET: BASELINE | TOTAL FREQUENCY (N)= 777 |
| | OBSERVATIONS (OBS)= 777 |

ANALYSIS OF VARIANCE TABLE

| SOURCE | DF | CHI-SQUARE | PROB |
|---|---|---|---|
| INTERCEPT | 1 | 95.46 | 0.0001 |
| PARCLOG | 1 | 54.96 | 0.0001 |
| AGULHA | 1 | 10.28 | 0.0013 |
| RETOENTB | 1 | 36.74 | 0.0001 |
| MAOARTB | 1 | 4.28 | 0.0385 |
| PARCLOG*MAOARTB | 1 | 1.29 | 0.2553 |
| LIKELIHOOD RATIO | 245 | 265.69 | 0.1737 |

ANALYSIS OF INDIVIDUAL PARAMETERS

| EFFECT | PARAMETER | ESTIMATE | STANDARD ERROR | CHI-SQUARE | PROB |
|---|---|---|---|---|---|
| INTERCEPT | 1 | 2.74564 | 0.281018 | 95.46 | 0.0001 |
| PARCLOG | 2 | -.568652 | .0767078 | 54.96 | 0.0001 |
| AGULHA | 3 | -1.22404 | 0.381823 | 10.28 | 0.0013 |
| RETOENTB | 4 | -1.21765 | 0.200884 | 36.74 | 0.0001 |
| MAOARTB | 5 | -1.0308 | 0.497986 | 4.28 | 0.0385 |
| PARCLOG*MAOARTB | 6 | 0.174129 | 0.153078 | 1.29 | 0.2553 |

UNIVARIATE PROCEDURE

Variable=GI

Moments

| | | | |
|---|---|---|---|
| N | 251 | Sum Wgts | 251 |
| Mean | 0.066333 | Sum | 16.64947 |
| Std Dev | 1.031218 | Variance | 1.063411 |
| Skewness | -0.57852 | Kurtosis | -0.44905 |
| USS | 266.957 | CSS | 265.8526 |
| CV | 1554.618 | Std Mean | 0.06509 |
| T:Mean=0 | 1.019092 | Prob>\|T\| | 0.3091 |
| Sgn Rank | 1613 | Prob>\|S\| | 0.1617 |
| Num ^= 0 | 251 | | |

Quantiles(Def=5)

| | | | |
|---|---|---|---|
| 100% Max | 2.327207 | 99% | 1.941333 |
| 75% Q3 | 0.801867 | 95% | 1.398005 |
| 50% Med | 0.353579 | 90% | 1.158646 |
| 25% Q1 | -0.87676 | 10% | -1.48755 |
| 0% Min | -2.89117 | 5% | -1.72546 |
| | | 1% | -2.28492 |
| Range | 5.218376 | | |
| Q3-Q1 | 1.678628 | | |
| Mode | 0.351048 | | |

Extremes

| Lowest | Obs | Highest | Obs |
|---|---|---|---|
| -2.89117( | 245) | 1.738377( | 25) |
| -2.87306( | 113) | 1.779321( | 68) |
| -2.28492( | 138) | 1.941333( | 14) |
| -2.15783( | 248) | 2.222968( | 103) |
| -2.08573( | 177) | 2.327207( | 93) |

```
                Histogram                  #

 2.25+*                                    2
     .*****                                9
     .***************                     27
     .*********************************   65
     .****************************        56
-0.25+**********                          18
     .************                        22
     .***************                     28
     .**********                          18
     .**                                   4
-2.75+*                                    2
      ----+----+----+----+----+----+---
      * may represent up to 2 counts
```

```
                Normal Probability Plot

 2.25+                                                    +++++ *
     |                                                 ++++++++++
     |                                            ********
     |                                       *********
     |                                  ********++
-0.25+                             *****++
     |                        ++*****
     |                   ++********
     |              ++********
     |         +*****
-2.75+**
     +----+----+----+----+----+----+----+----+----+----+
         -2        -1         0         +1        +2
```

UNIVARIATE PROCEDURE

Variable=XI

Moments

| | | | |
|---|---|---|---|
| N | 251 | Sum Wgts | 251 |
| Mean | 0.012093 | Sum | 3.035445 |
| Std Dev | 0.987407 | Variance | 0.974972 |
| Skewness | -0.818 | Kurtosis | 1.106168 |
| USS | 243.7796 | CSS | 243.7429 |
| CV | 8164.834 | Std Mean | 0.062325 |
| T:Mean=0 | 0.194039 | Prob>\|T\| | 0.8463 |
| Sgn Rank | 1293 | Prob>\|S\| | 0.2623 |
| Num ^= 0 | 251 | | |

Quantiles(Def=5)

| | | | |
|---|---|---|---|
| 100% Max | 2.401082 | 99% | 1.879137 |
| 75% Q3 | 0.647919 | 95% | 1.287257 |
| 50% Med | 0.290381 | 90% | 0.978081 |
| 25% Q1 | -0.71614 | 10% | -1.37053 |
| 0% Min | -3.76367 | 5% | -1.69642 |
| | | 1% | -3.04279 |
| Range | 6.164753 | | |
| Q3-Q1 | 1.364055 | | |
| Mode | 0.252102 | | |

Extremes

| Lowest | Obs | Highest | Obs |
|---|---|---|---|
| -3.76367( | 245) | 1.820204( | 88) |
| -3.70793( | 113) | 1.849958( | 93) |
| -3.04279( | 248) | 1.879137( | 25) |
| -2.79345( | 177) | 2.337994( | 14) |
| -2.31882( | 138) | 2.401082( | 103) |

UNIVARIATE PROCEDURE

Variable=HII

Moments

| | | | |
|---|---|---|---|
| N | 251 | Sum Wgts | 251 |
| Mean | 0.01992 | Sum | 5 |
| Std Dev | 0.018487 | Variance | 0.000342 |
| Skewness | 2.5646 | Kurtosis | 10.65042 |
| USS | 0.185043 | CSS | 0.085442 |
| CV | 92.80433 | Std Mean | 0.001167 |
| T:Mean=0 | 17.07138 | Prob>\|T\| | 0.0001 |
| Sgn Rank | 15813 | Prob>\|S\| | 0.0001 |
| Num ^= 0 | 251 | | |

Quantiles(Def=5)

| | | | |
|---|---|---|---|
| 100% Max | 0.147375 | 99% | 0.081336 |
| 75% Q3 | 0.027384 | 95% | 0.056337 |
| 50% Med | 0.013714 | 90% | 0.041794 |
| 25% Q1 | 0.007365 | 10% | 0.0061 |
| 0% Min | 0.002901 | 5% | 0.004231 |
| | | 1% | 0.002906 |
| Range | 0.144475 | | |
| Q3-Q1 | 0.020018 | | |
| Mode | 0.007672 | | |

Extremes

| Lowest | Obs | Highest | Obs |
|---|---|---|---|
| 0.002901( | 85) | 0.068723( | 67) |
| 0.002906( | 97) | 0.07291( | 21) |
| 0.002906( | 96) | 0.081336( | 157) |
| 0.002917( | 82) | 0.107042( | 194) |
| 0.002945( | 75) | 0.147375( | 5) |

UNIVARIATE PROCEDURE

Variable=C1I

Moments

| | | | |
|---|---|---|---|
| N | 251 | Sum Wgts | 251 |
| Mean | 0.024093 | Sum | 6.047262 |
| Std Dev | 0.063031 | Variance | 0.003973 |
| Skewness | 5.605067 | Kurtosis | 37.76948 |
| USS | 1.138908 | CSS | 0.993213 |
| CV | 261.6171 | Std Mean | 0.003978 |
| T:Mean=0 | 6.055788 | Prob>\|T\| | 0.0001 |
| Sgn Rank | 15813 | Prob>\|S\| | 0.0001 |
| Num ^= 0 | 251 | | |

Quantiles(Def=5)

| | | | |
|---|---|---|---|
| 100% Max | 0.56747 | 99% | 0.377374 |
| 75% Q3 | 0.018883 | 95% | 0.11015 |
| 50% Med | 0.00558 | 90% | 0.050035 |
| 25% Q1 | 0.00139 | 10% | 0.000578 |
| 0% Min | 4.169E-7 | 5% | 0.000293 |
| | | 1% | 0.000027 |
| Range | 0.56747 | | |
| Q3-Q1 | 0.017493 | | |
| Mode | 0.000578 | | |

Extremes

| Lowest | Obs | Highest | Obs |
|---|---|---|---|
| 4.169E-7( | 89) | 0.225534( | 245) |
| 0.000019( | 38) | 0.290574( | 21) |
| 0.000027( | 28) | 0.377374( | 103) |
| 0.000055( | 64) | 0.484915( | 113) |
| 0.000074( | 251) | 0.56747( | 5) |